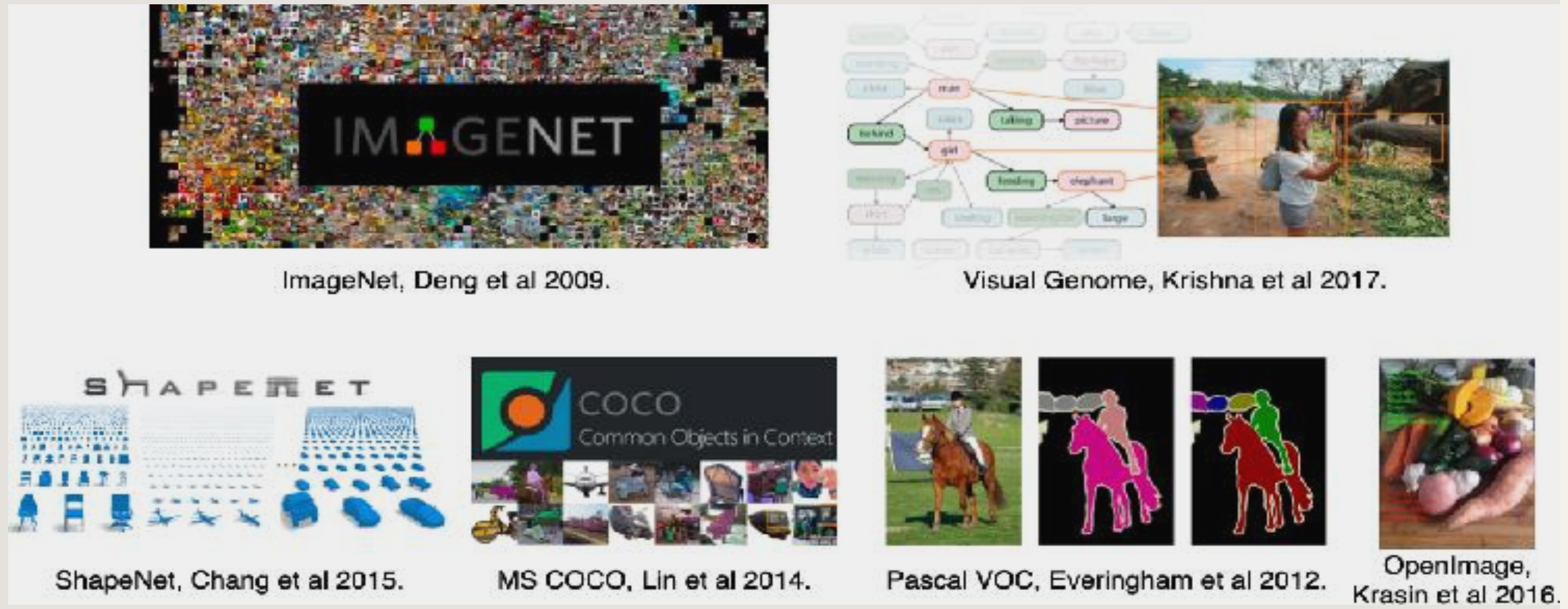
Nicolas Bosteels

Embodied Reasoning Through Planning with Language and Vision Foundation Models

Datasets



ImageNet, Deng et al 2009.

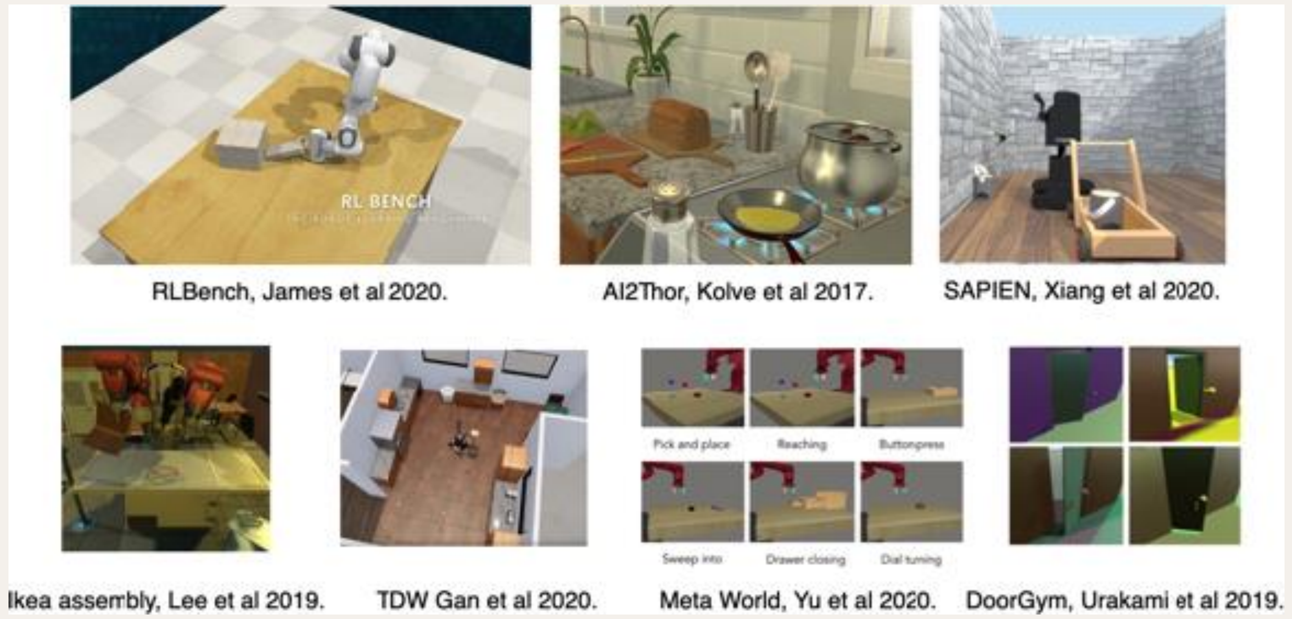
Visual Genome, Krishna et al 2017.

ShapeNet, Chang et al 2015.

MS COCO, Lin et al 2014.

Pascal VOC, Everingham et al 2012.

OpenImage, Krasin et al 2016.



RLBench, James et al 2020.

AI2Thor, Kolve et al 2017.

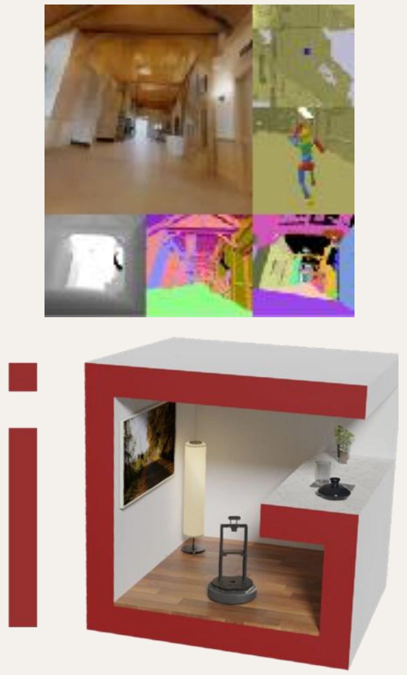
SAPIEN, Xiang et al 2020.

Ikea assembly, Lee et al 2019.

TDW Gan et al 2020.

Meta World, Yu et al 2020.

DoorGym, Urakami et al 2019.



Tasks

Classification

Segmentation

Detection

Generation

Captioning

...

Visual Navigation

Manipulation

Rearragement

Embodied-QA

Mobile Manipulation

Instruction Following

...

Internet AI

Embodied AI

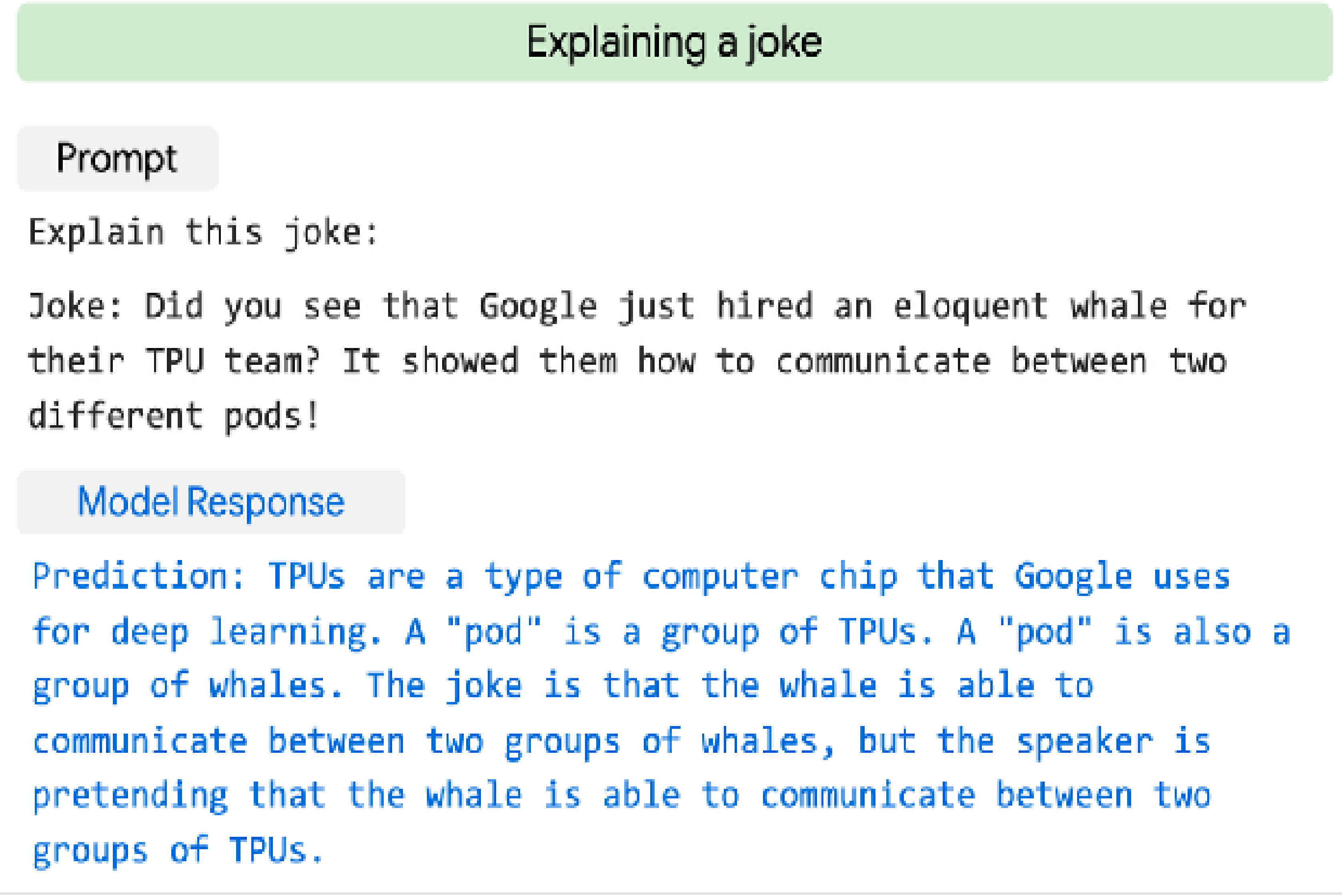
Motivation

For a robot to intelligently act in the world it needs:

- Decompose long-horizon tasks into a sequence of simpler tasks.
- Have a large repertoire of these simpler tasks and ability to learn these online.
- Perceptual feedback to recognize and change plans when conditions change or actions fail.
- World knowledge so we (lazy humans) can spend as little effort as possible embedding prior knowledge into the model.
- Perform these tasks in real-time, low latency needed.

Mixing language and robotics: Large Language Models (LLMs)

- Contain rich internalized knowledge about the world.
- Can interpret natural language instructions, code, mathematics…
- Can handle large sequences of complex data.
- Perform high-level reasoning about a multitude of diverse topics.
- Can even help me make this presentation.



Explaining a joke

Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Can LLMs serve as reasoning models that combine multiple sources of feedback and become interactive problem solvers for embodied tasks?

Mixing language and robotics: Large Language Models (LLMs)

Challenges:

1. Robot Language: Our robots can only do a fixed number of commands and need the problem broken down in actionable steps. This is not what LLMs have seen.

2. Grounding: LLMs have not directly “experienced” the physical world.

3. Safety, alignment, interpretability…

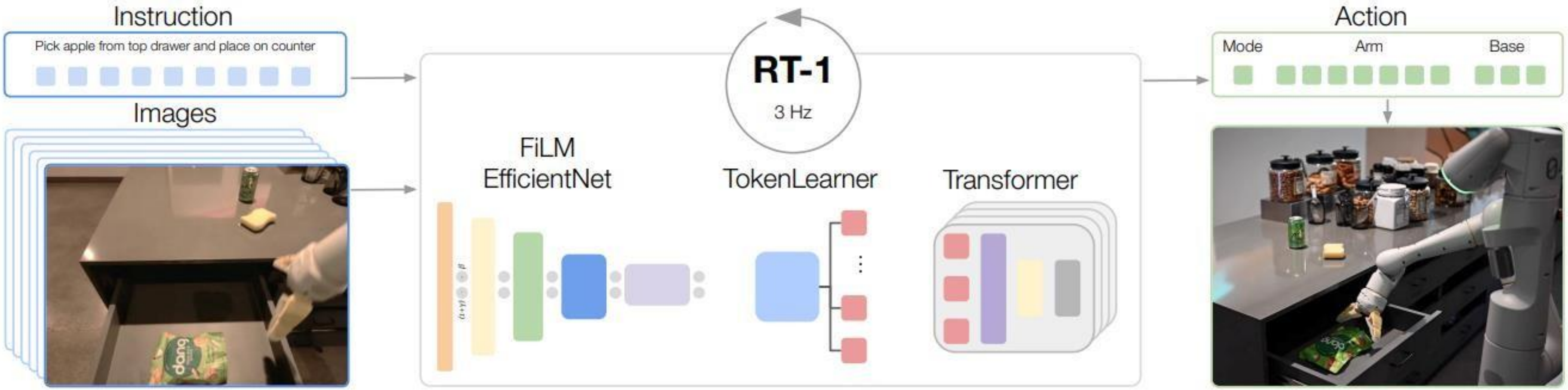
I spilled my drink, can you help?

I’m feeling tired, can you make me a pumpkin spiced latte?

RT-1: Robotics Transformer for Real-World Control at Scale

- Behavior Cloning:
  - Input: Observations (Camera, Prior-perception...) and Task Description
  - Output: Action (e.g. timesteps x join_states)

- Training Data:
  - Human demonstrations (Language, Observation, Action)-pairs as Supervision



Language Conditioned Robot Behavior

- Naive language conditioned imitation learning works on short horizon tasks but struggles with long-horizon tasks and complex instructions.

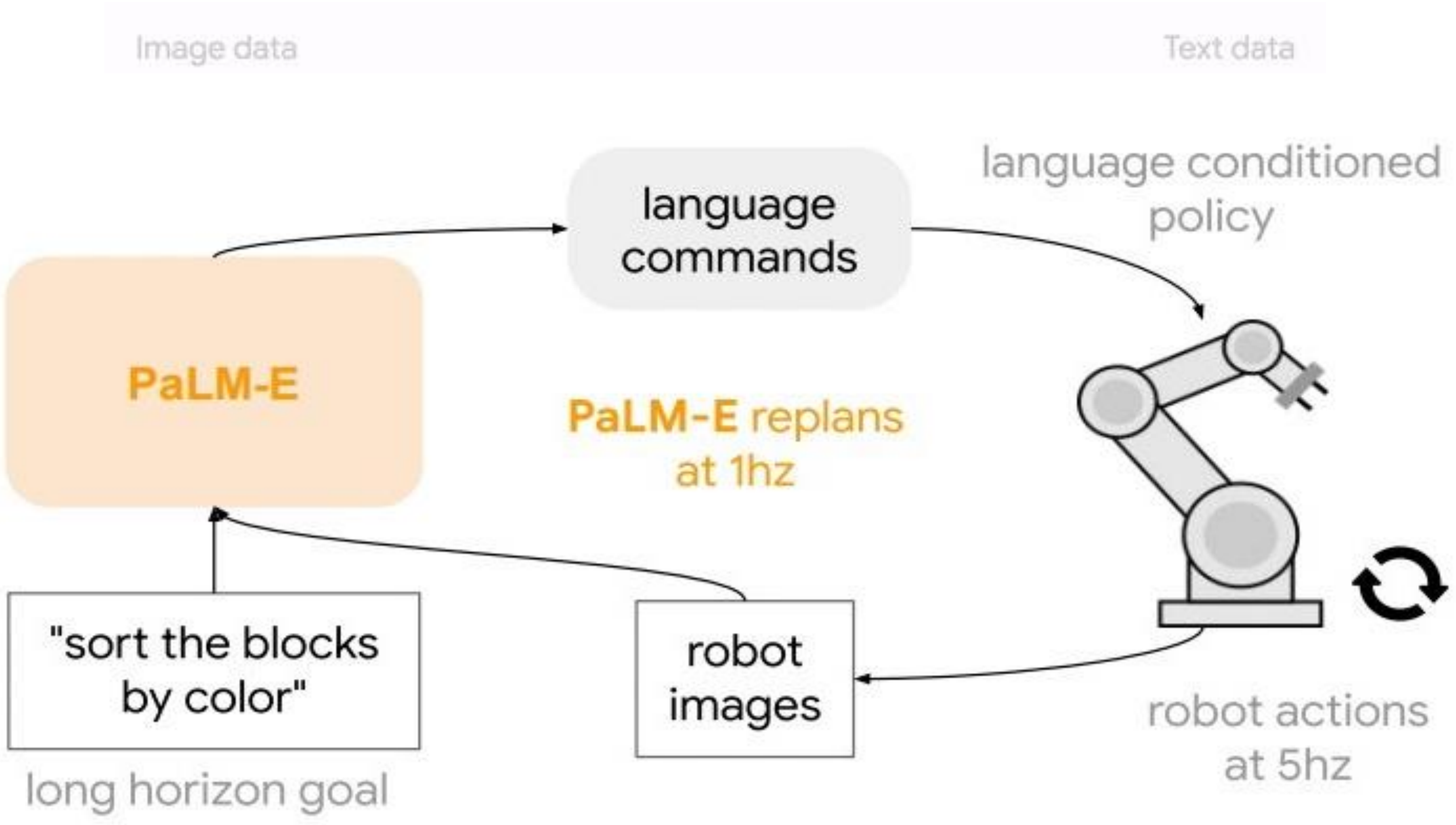
- Fail to generalize to out of distribution tasks not seen during training.

"I just worked out, can you bring me a snack and a drink to recover?"

"I'm feeling tired, can you bring me the key to happiness?"

Fails to resolve to grab a beer due to a lack of semantic and world knowledge.

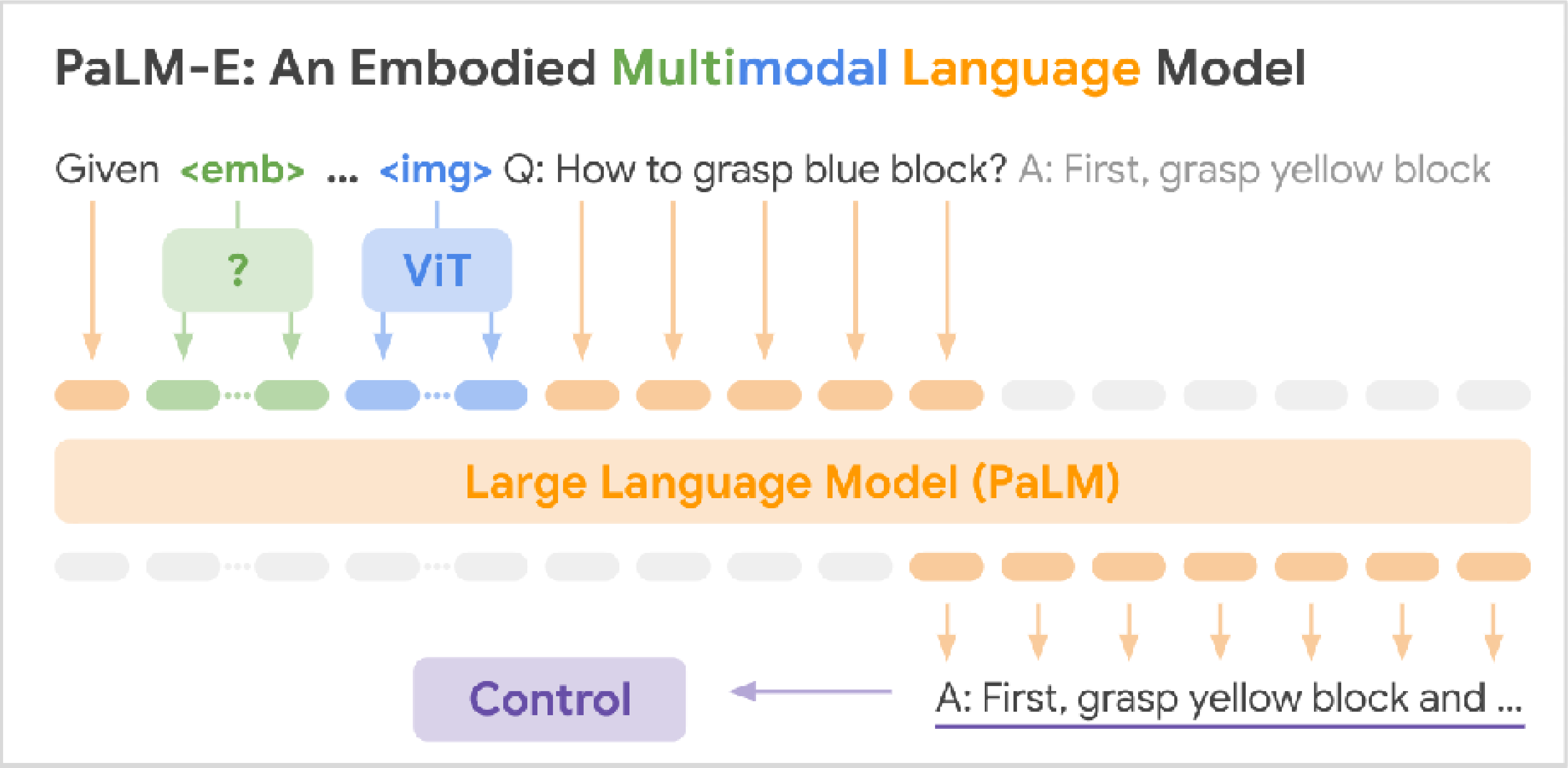
PaLM-E: An Embodied Multimodal Language Model

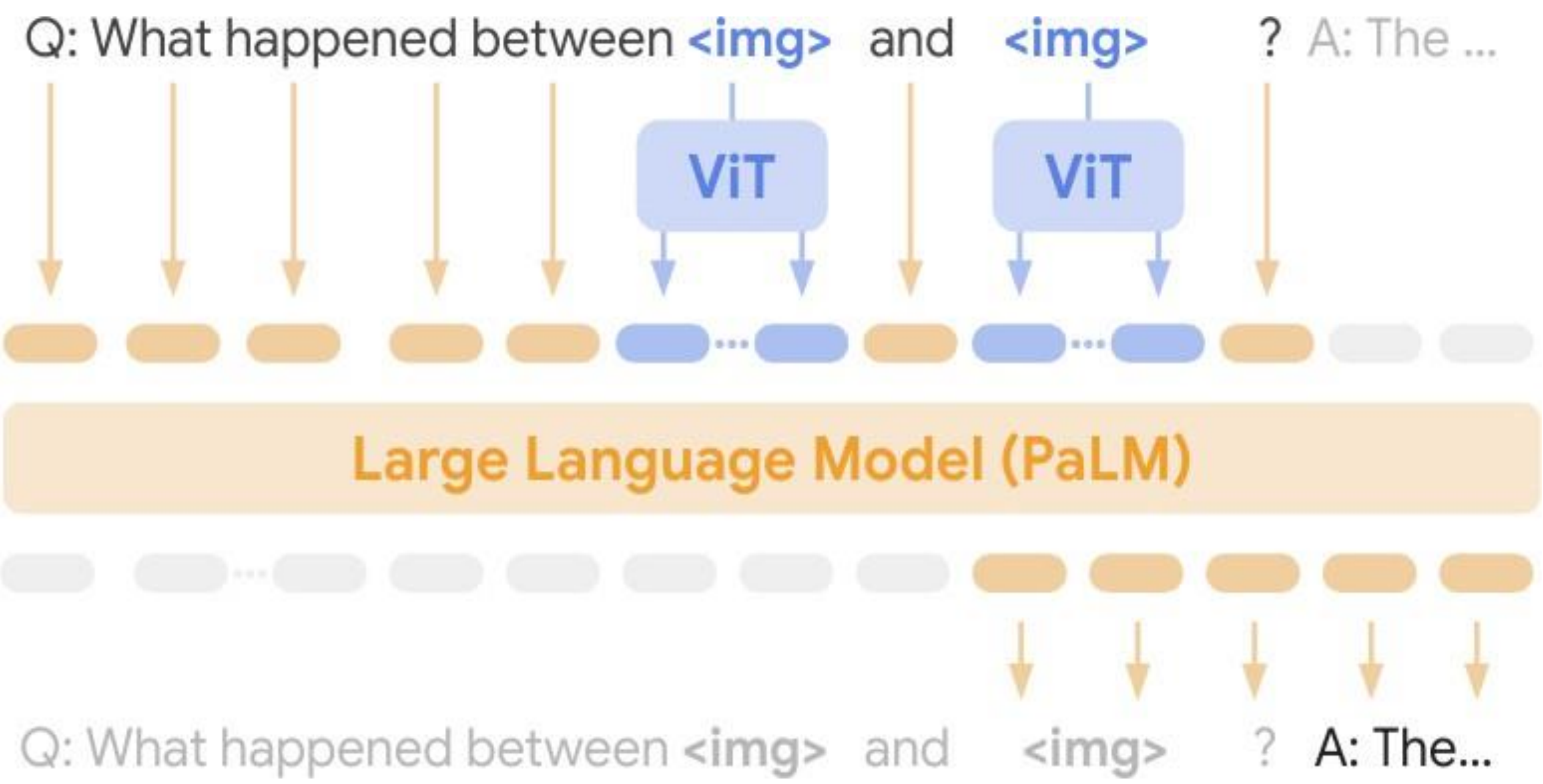


Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, Pete Florence

Google Research

Architecture of PaLM-E


PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and ...


Q: What happened between <img> and <img> ? A: The ...
ViT
ViT
Large Language Model (PaLM)
Q: What happened between <img> and <img> ? A: The...

Arbitrary interleaving

- Original token embedding space of LLM:

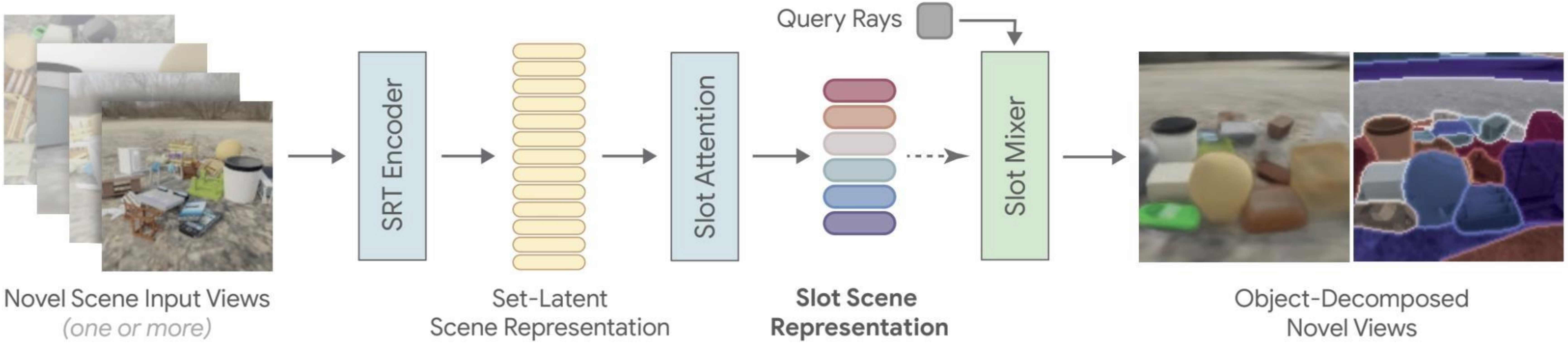
  $k \times W$ (k is the dim of embedding, W is the vocabulary)

  $W = 256000$

- How to *enable multi-modal inputs?*
  - *Project inputs into LLM token embedding space*

Approaches to Encoding Multi-Modal Features

- How to project image inputs into LLM embedding space? Choices:
  - State Estimation Vectors
  - ViT pretrained on image-classification
  - Object-centric ViT (mask + ViT)
  - Object Scene Representation Transformer (OSRT)



“Object Scene Representation Transformer” NeurIPS 2022

Approaches to Encoding Multi-Modal Features

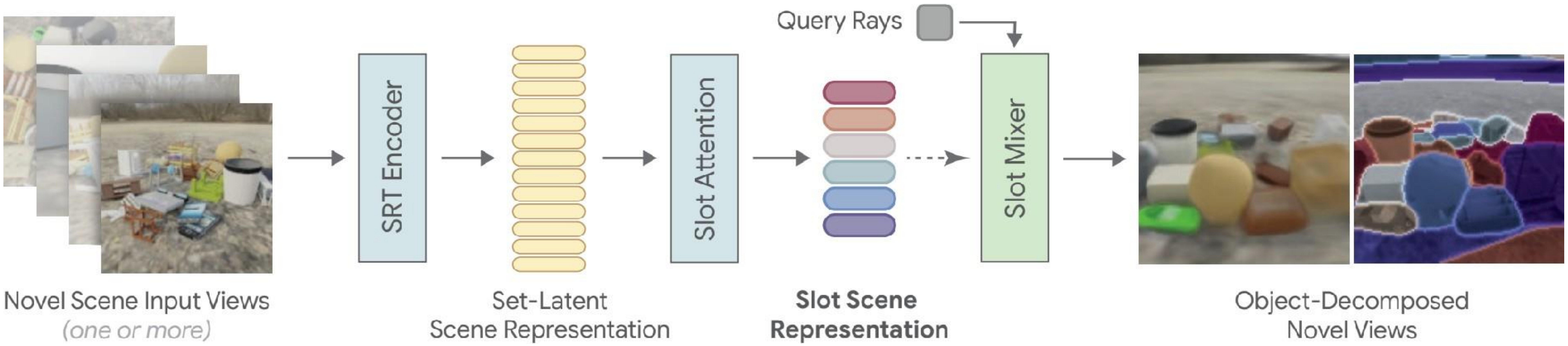
State Estimation Vectors

- State vectors from a robot or a state estimate for objects.
- Could be pose, color, size…
- An MLP layer projects these into one token (same dimensionality as text).
- Compact descriptions of the scene.

ViT

- Transformer architecture mapping an image into multiple token embeddings or patches.
- These embeddings are not the same dimensionality as the language tokens.
- For each embedding, learn a transformation that projects it to the language dimensionality.

Scene Representation: Object Scene Representation Transformer

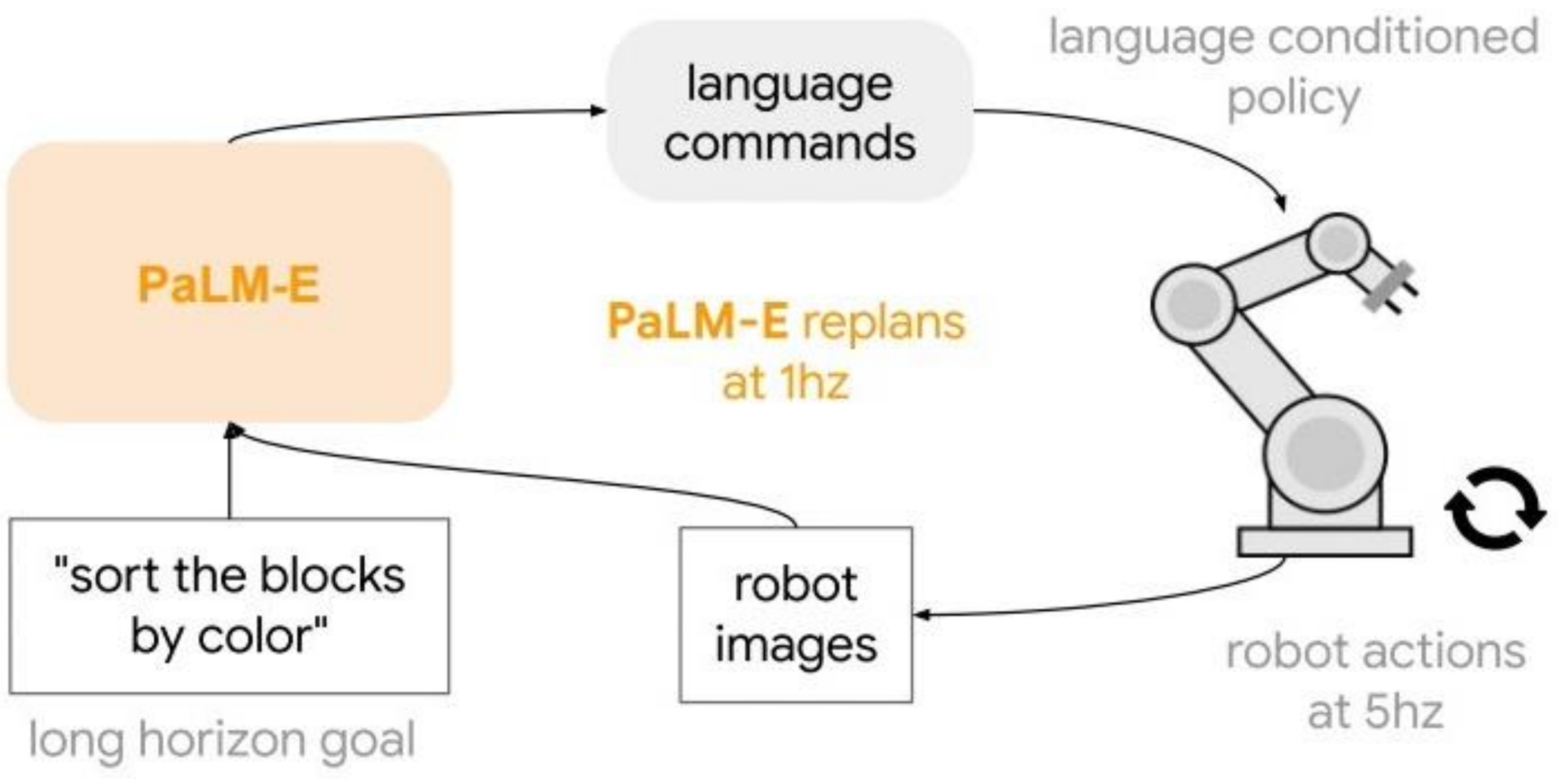


- Takes multiple 2D views.
- Reconstructs a latent 3D space.
- Decomposes the space into objects.
- Outputs object tokens with 3D structure.

- Unsupervised learning of 3D neural scene representations.
- Ideal for spatial reasoning but more expensive to train.

$o_j = \bar{\phi}_{\mathrm{OSRT}}(I_{1:v})_j \in \mathbb{R}^{\bar{k}}$ $\quad$ $x^j_{1:m} = \psi(\bar{\phi}_{\mathrm{OSRT}}(I_{1:v})_j)$ with an MLP ψ.

Training PaLM-E

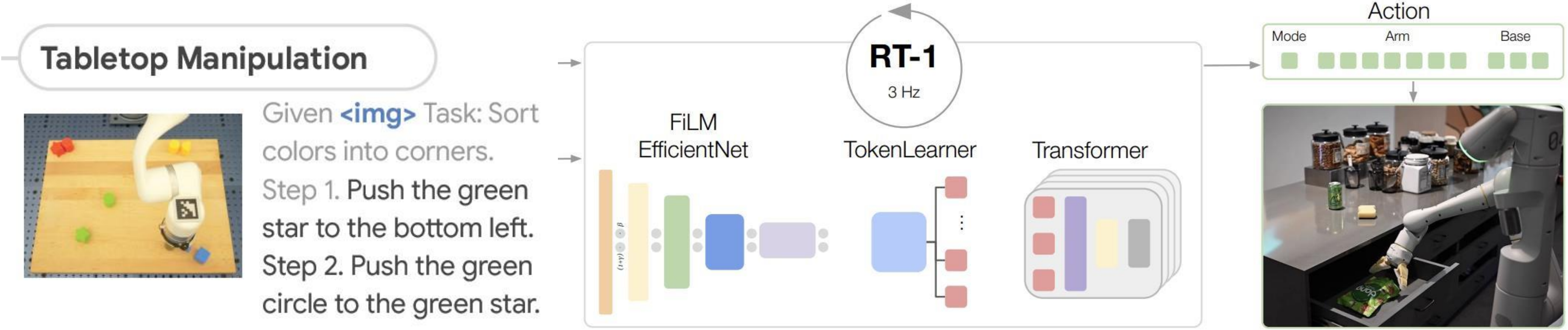


- Training:
  - Prefix-Decoder-only LLM
    - Prefix: $w_{1:n}$ (image, states, queries)
  - Cross-entropy loss over the next-token prediction on text outputs: $w_{n+1:L}$

$$p(w_{n+1:L}|w_{1:n}) = \prod_{l=n+1}^{L} p_{\mathrm{LM}}(w_l|w_{1:l-1}).$$

RT-1 + PaLM-E

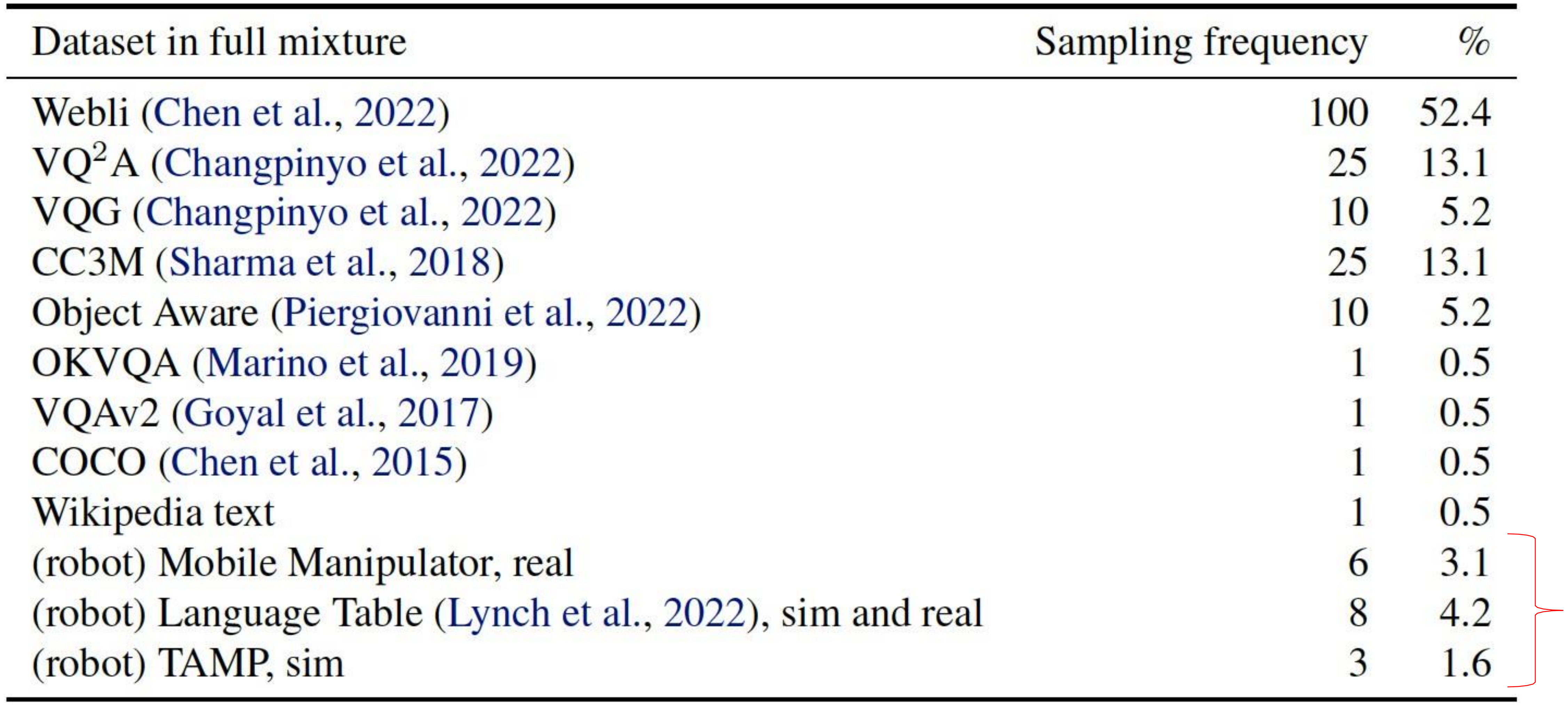
- Output: languages describing how to complete the task
- Use language guided low-level control to control the robots (RT-1)
  - Palm-e outputs language guidance
  - RT-1 map the language to low-level actions



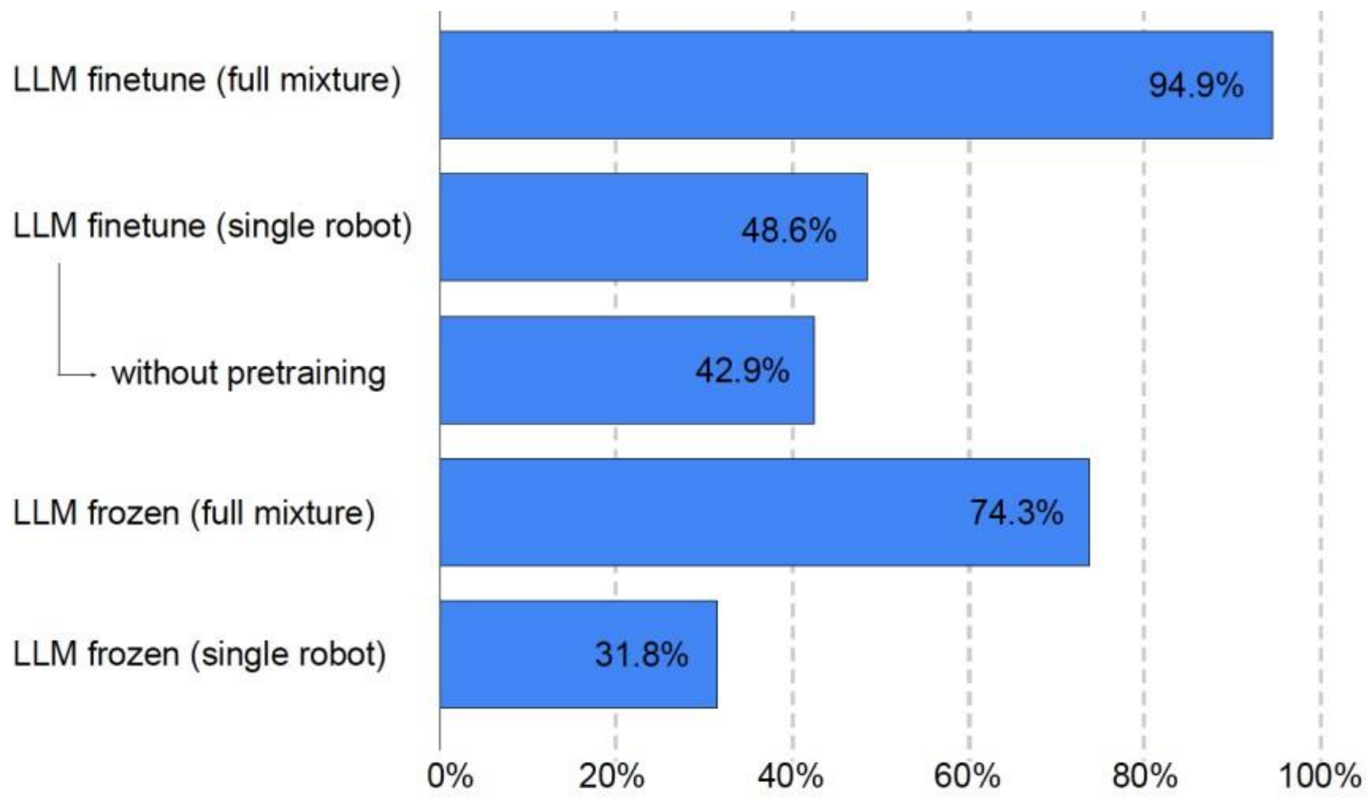
Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

Dataset and Training

- Training:
  - LLM + Visual Encoder
    Different model-freezing strategies: (LLM, visual encoder, projector)
    - Fully fine-tuning
    - Freeze LLM, fine-tune encoder and projector only
  - Co-training across tasks:
    - Mixture of datasets of different tasks



Dataset in full mixture	Sampling frequency	%
Webli (Chen et al., 2022)	100	52.4
VQ^{2}A (Changpinyo et al., 2022)	25	13.1
VQG (Changpinyo et al., 2022)	10	5.2
CC3M (Sharma et al., 2018)	25	13.1
Object Aware (Piergiovanni et al., 2022)	10	5.2
OKVQA (Marino et al., 2019)	1	0.5
VQAv2 (Goyal et al., 2017)	1	0.5
COCO (Chen et al., 2015)	1	0.5
Wikipedia text	1	0.5
(robot) Mobile Manipulator, real	6	3.1
(robot) Language Table (Lynch et al., 2022), sim and real	8	4.2
(robot) TAMP, sim	3	1.6

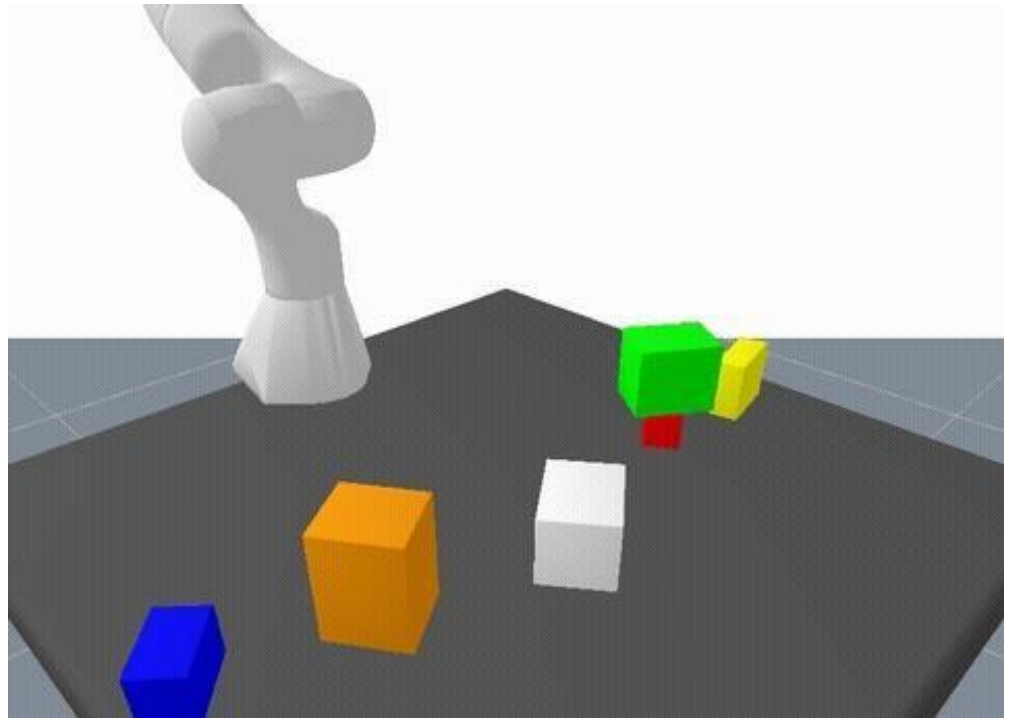


Embodied data: 8.9%

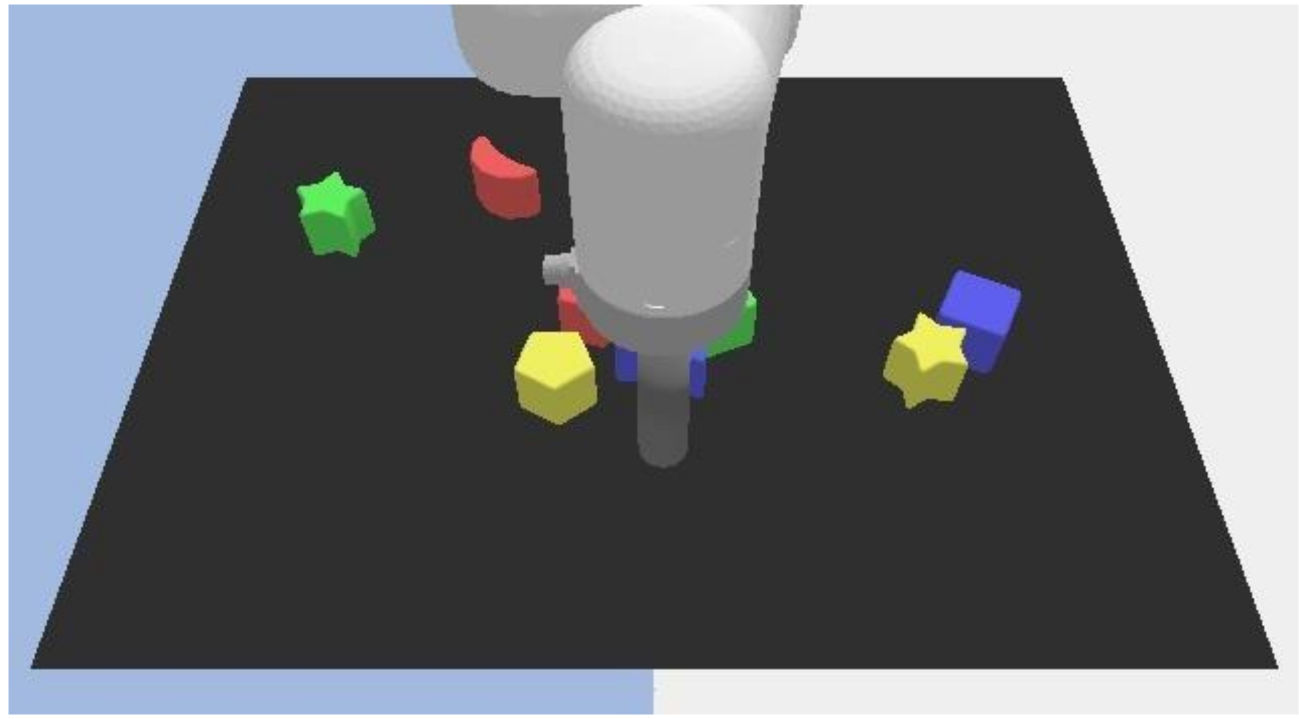
Experiments and Results

Consists of diverse robotic mobile manipulation tasks across

Robot Environments



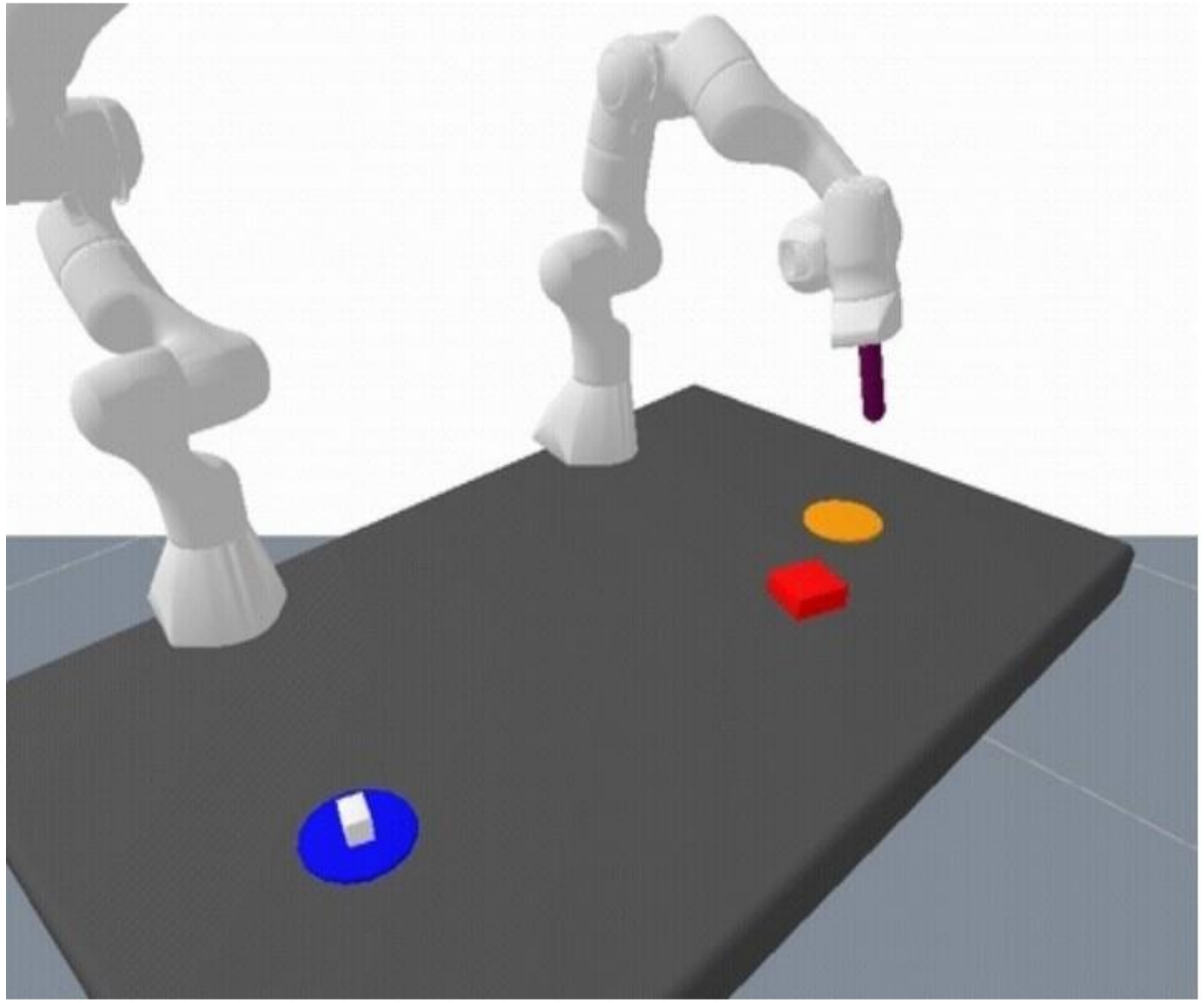
Task and Motion Planning (TAMP)

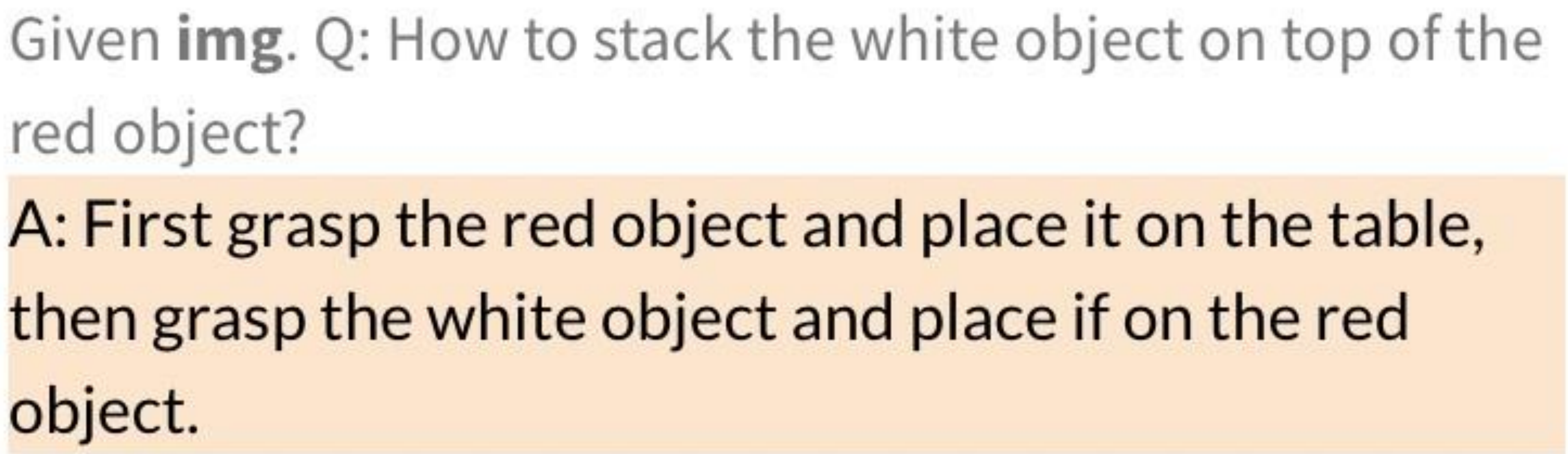


Language Table

Mobile manipulation (SayCan)

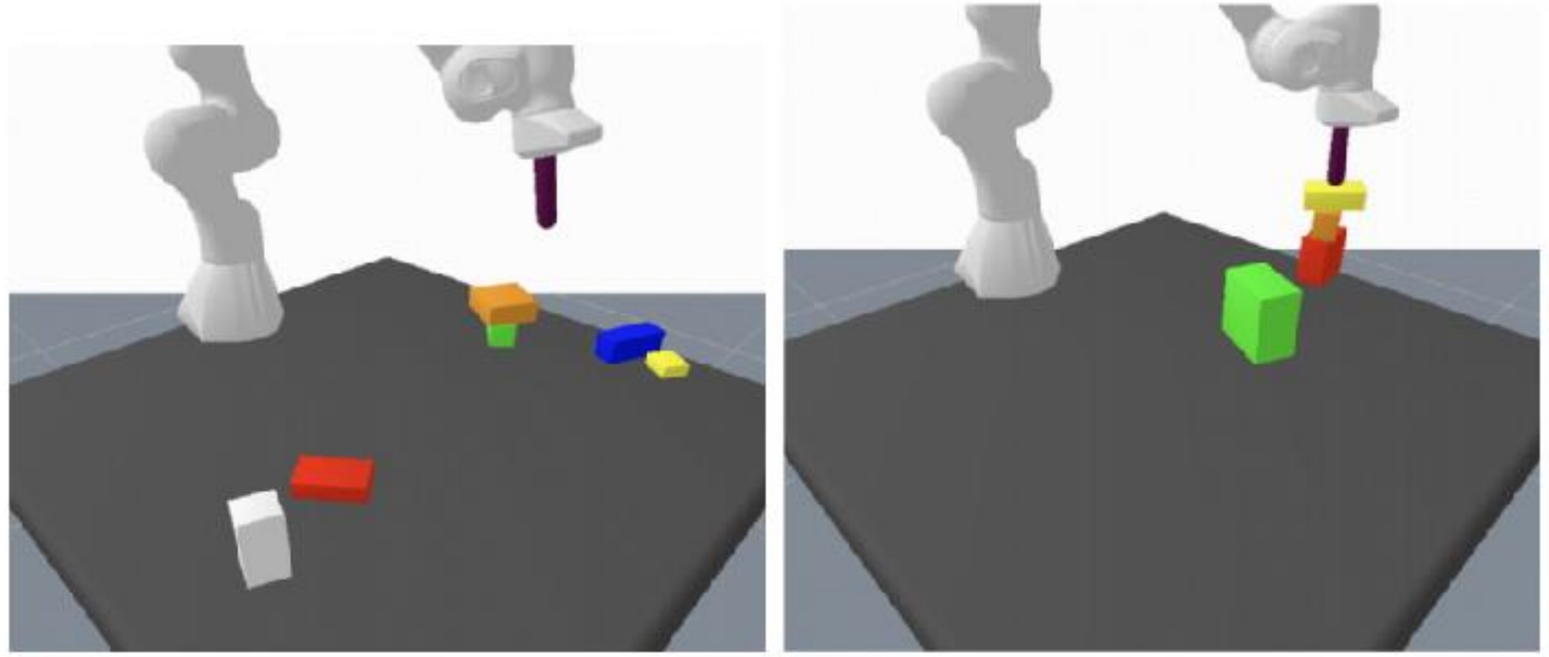
PaLM-E on TAMP-like tasks





Given **img**. Q: How to stack the white object on top of the red object?

A: First grasp the red object and place it on the table, then grasp the white object and place if on the red object.



Experiments and Results

Task and Motion Planning (TAMP) Environment

- Robot has to manipulate (grasp and stack) objects.
- Training scenes contain 3-5 cube-shaped objects of different sizes, colors and samples initial poses

VQA Tasks

- **q1: color of an object**
- **q2: object-table relation.**
  Example prompt:
  Given <img>. Q: *Is the red object left, right, or center of the table?*
  Target: A: *The red object is in the center of the table.*
- **q3: object-object relations.**
  Example prompt:
  Given <img>. Q: *Is the yellow object below the blue object?*
  Target: A: *No, the yellow object is not below the blue object.*
- **q4: plan feasibility.**
  Example prompt:
  Given <img>. Q: Is it possible to first grasp the blue object, pace it on the yellow object, and then grasp the yellow object?.
  Target: A: *No, this is not possible.*

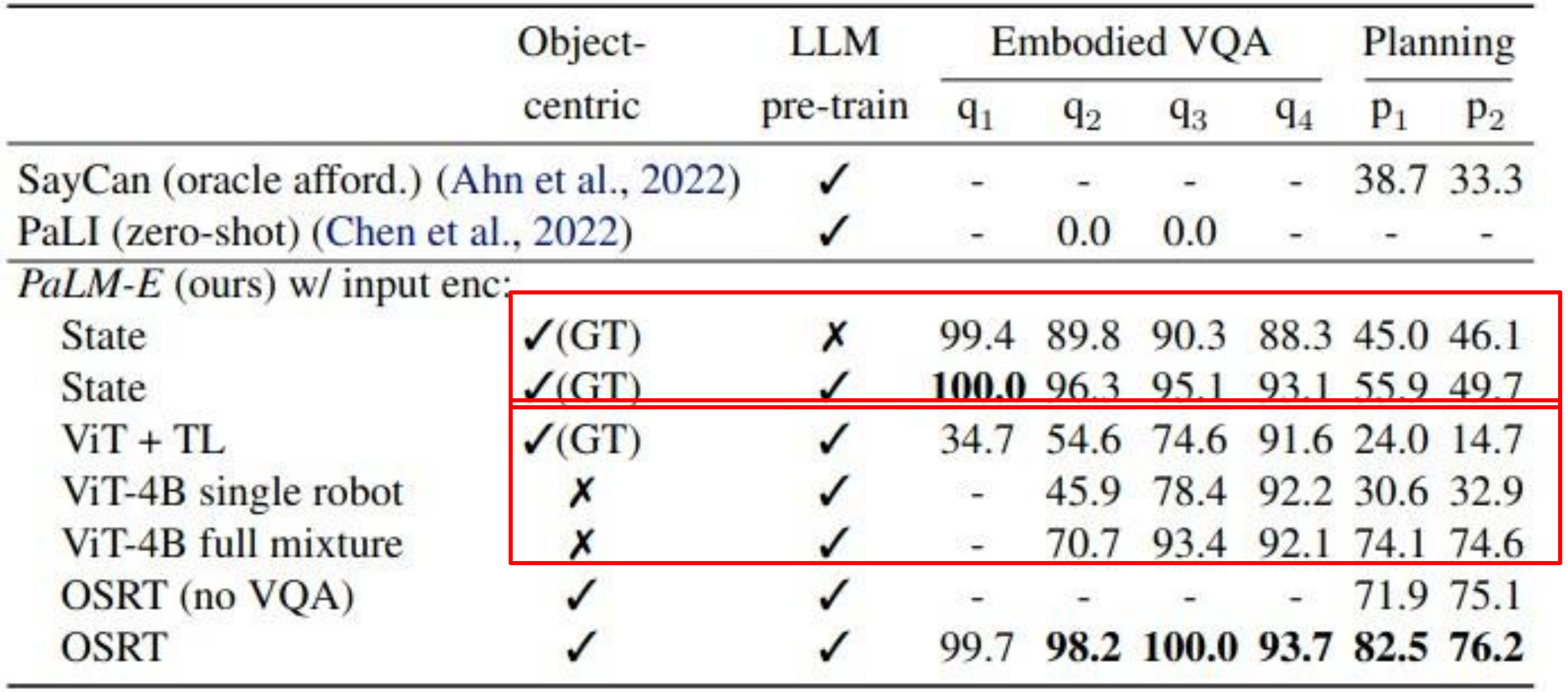
Planning Tasks

- **p1: grasping.**
  Example prompt:
  Given <img>. Q: *How to grasp the green object?.*
  Target: A: *First grasp the orange object and place it on the table, then grasp the green object.*
- **p2: stacking.**
  Example prompt:
  Given <img>. Q: How to stack the white object on top of the red object?.
  Target: A: First grasp the green object and place it on the table, then grasp the white object and place it on the red object.

Experiments and Results

TAMP Env

Goal: Compare different input configurations with respect to performance and data-efficiency

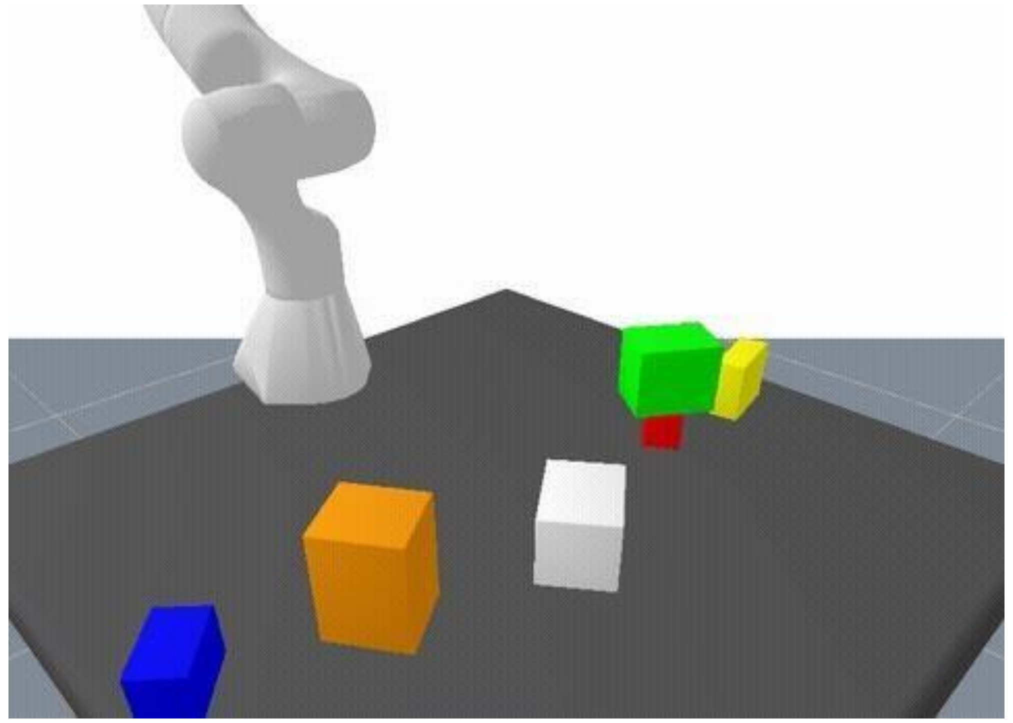


	Object-	LLM	Embodied VQA				Planning	
	centric	pre-train	q_1	q_2	q_3	q_4	p_1	p_2
SayCan (oracle afford.) (Ahn et al., 2022)		✓	-	-	-	-	38.7	33.3
PaLI (zero-shot) (Chen et al., 2022)		✓	-	0.0	0.0	-	-	-
PaLM-E (ours) w/ input enc:								
State	✓(GT)	✗	99.4	89.8	90.3	88.3	45.0	46.1
State	✓(GT)	✓	**100.0**	96.3	95.1	93.1	55.9	49.7
ViT + TL	✓(GT)	✓	34.7	54.6	74.6	91.6	24.0	14.7
ViT-4B single robot	✗	✓	-	45.9	78.4	92.2	30.6	32.9
ViT-4B full mixture	✗	✓	-	70.7	93.4	92.1	74.1	74.6
OSRT (no VQA)	✓	✓	-	-	-	-	71.9	75.1
OSRT	✓	✓	99.7	**98.2**	**100.0**	**93.7**	**82.5**	**76.2**

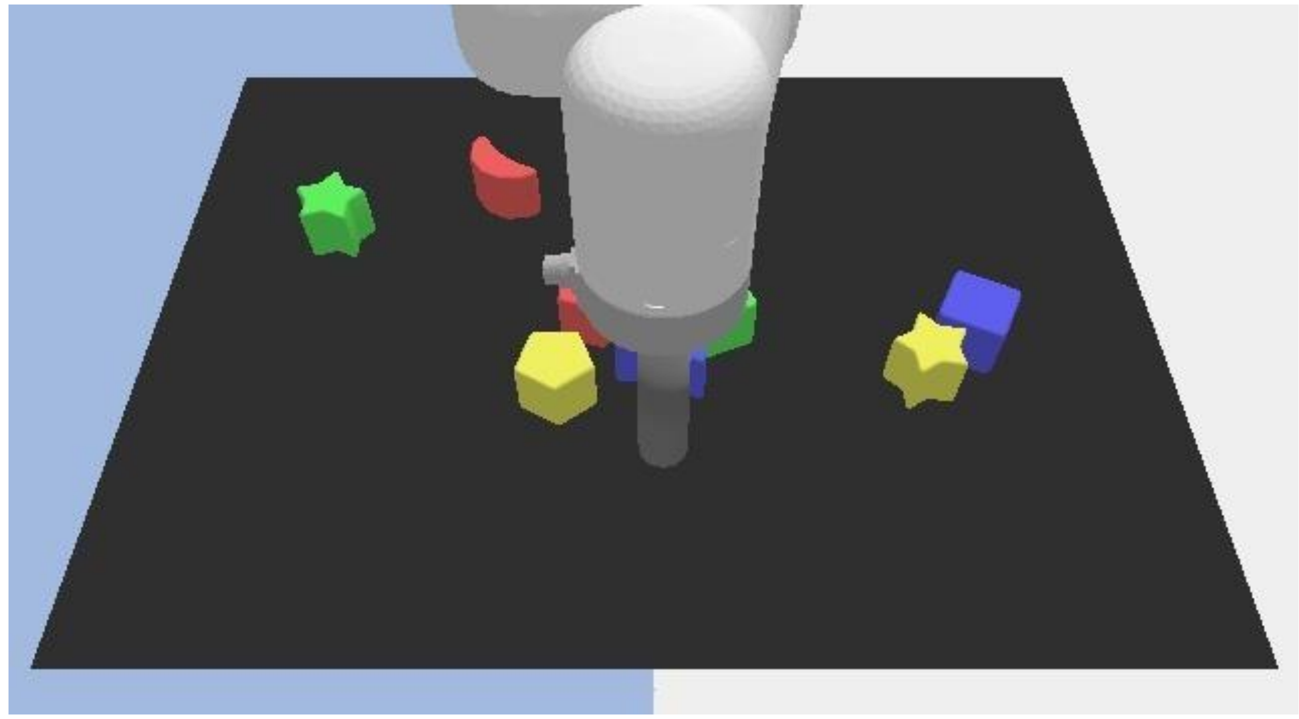
- Training on 1% of the TAMP dataset
- LLM is frozen, only the encoder and projector learn
- ViT variants: Effect of cross-domain transfer
- Using OSRT representation leads to best performance

Experiments and Results

Robot Environments



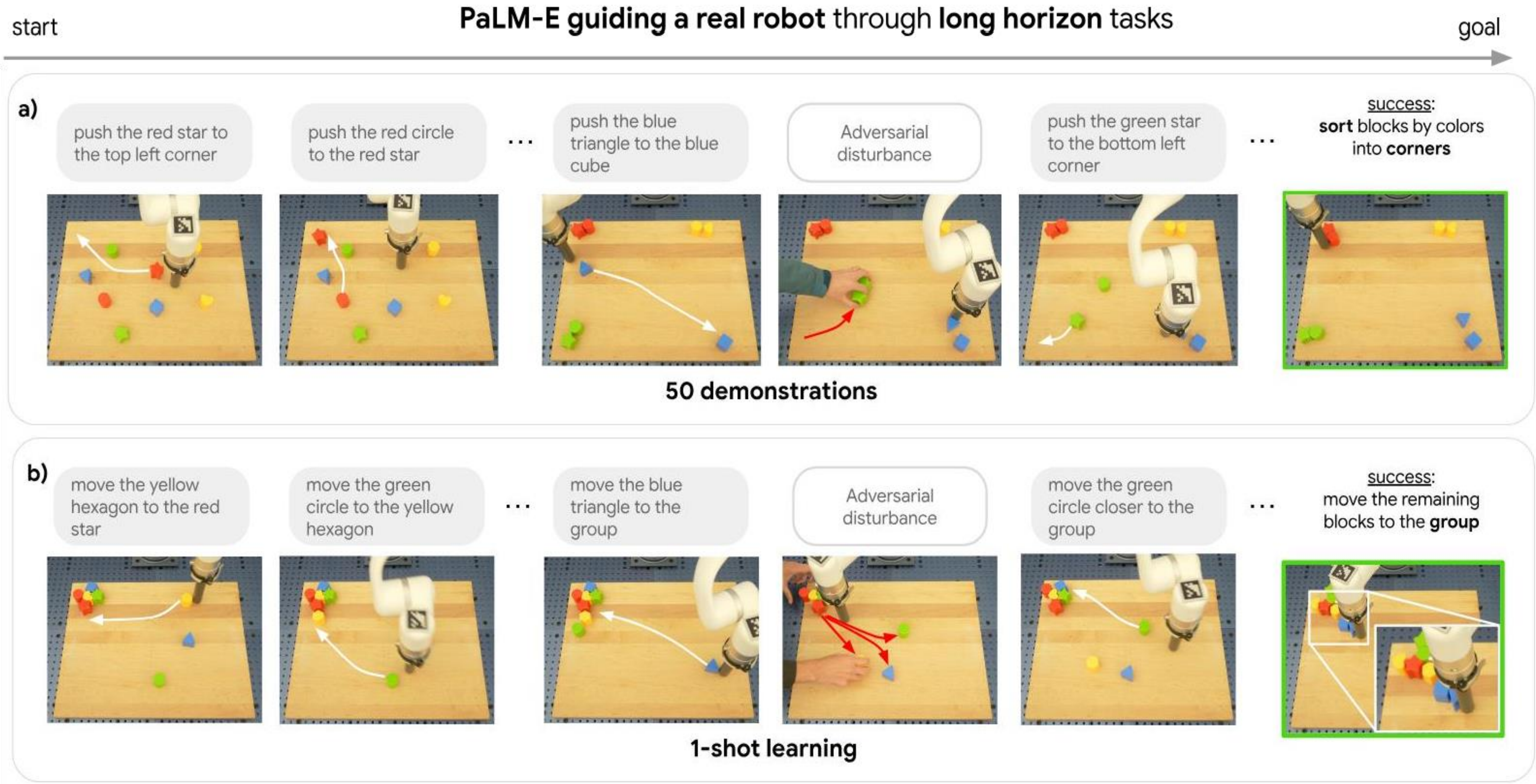
Task and Motion Planning (TAMP)



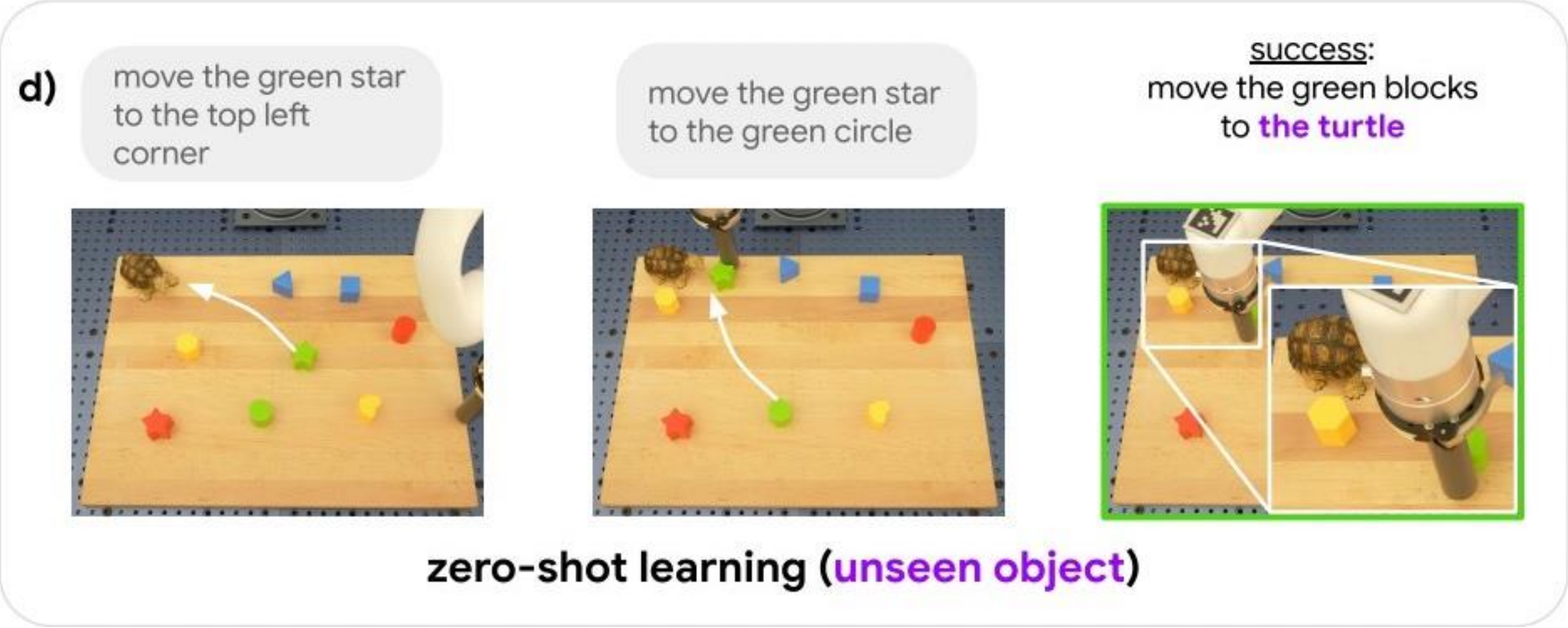
Language Table

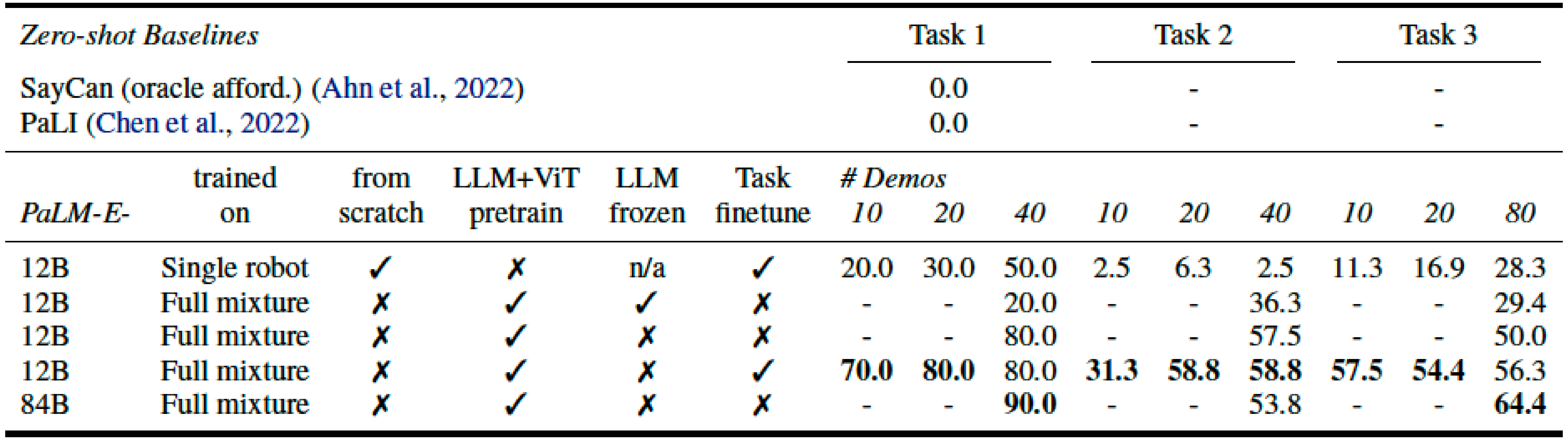
Mobile manipulation (SayCan)

Sample-efficient learning: Language Table Env


start
PaLM-E guiding a real robot through long horizon tasks
goal
a)
push the red star to the top left corner
push the red circle to the red star
...
push the blue triangle to the blue cube
Adversarial disturbance
push the green star to the bottom left corner
...
success:
sort blocks by colors into corners
50 demonstrations
b)
move the yellow hexagon to the red star
move the green circle to the yellow hexagon
...
move the blue triangle to the group
Adversarial disturbance
move the green circle closer to the group
...
success:
move the remaining blocks to the group
1-shot learning

Sample-efficient learning: Language Table Env

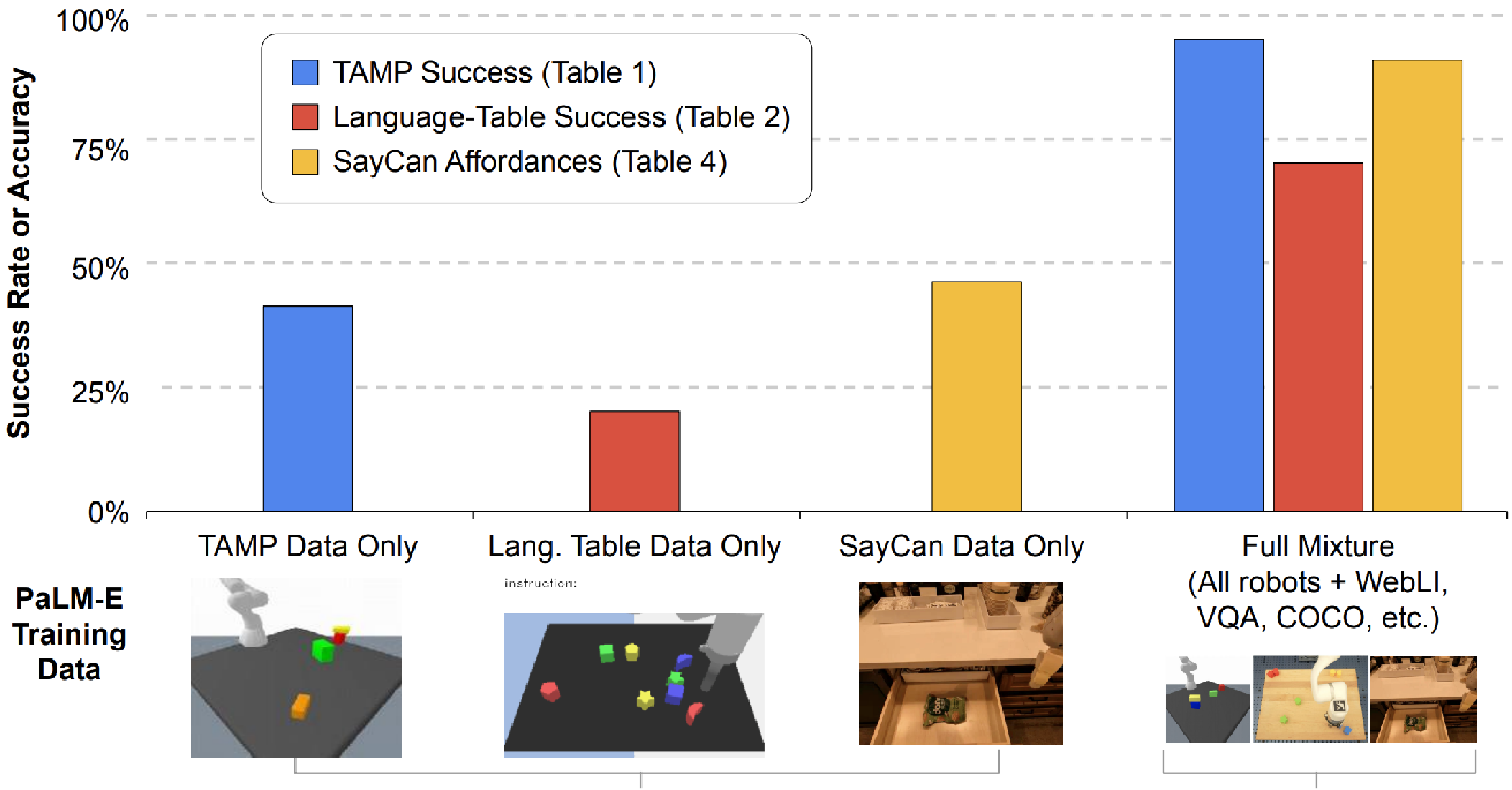

d)
move the green star to the top left corner
move the green star to the green circle
success:
move the green blocks to the turtle
zero-shot learning (unseen object)



Zero-shot Baselines							Task 1			Task 2			Task 3	
SayCan (oracle afford.) (Ahn et al., 2022)							0.0			-			-	
PaLI (Chen et al., 2022)							0.0			-			-	
	trained	from	LLM+ViT	LLM	Task	*# Demos*								
PaLM-E-	on	scratch	pretrain	frozen	finetune	*10*	*20*	*40*	*10*	*20*	*40*	*10*	*20*	*80*
12B	Single robot	✓	✗	n/a	✓	20.0	30.0	50.0	2.5	6.3	2.5	11.3	16.9	28.3
12B	Full mixture	✗	✓	✓	✗	-	-	20.0	-	-	36.3	-	-	29.4
12B	Full mixture	✗	✓	✗	✗	-	-	80.0	-	-	57.5	-	-	50.0
12B	Full mixture	✗	✓	✗	✓	**70.0**	**80.0**	80.0	**31.3**	**58.8**	**58.8**	**57.5**	**54.4**	56.3
84B	Full mixture	✗	✓	✗	✗	-	-	**90.0**	-	-	53.8	-	-	**64.4**

Table 2: Results on planning tasks in the simulated environment from Lynch et al. (2022).

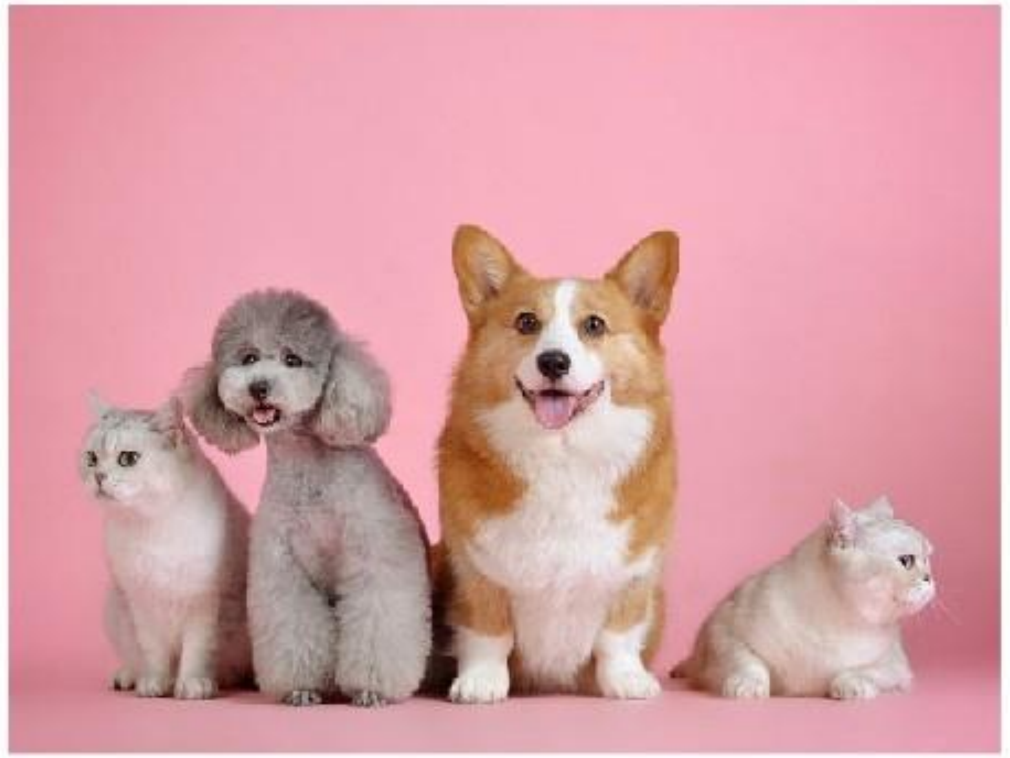
PaLM-E: Positive Transfer



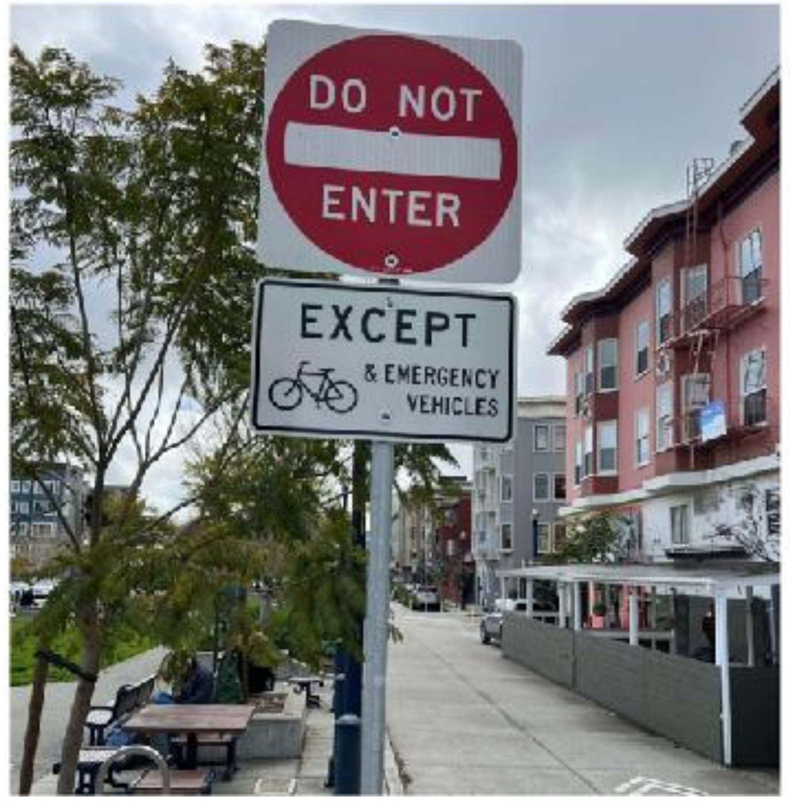
Different models for different robots, trained from scratch →

One model for all robots with ViT + PaLM pre-training

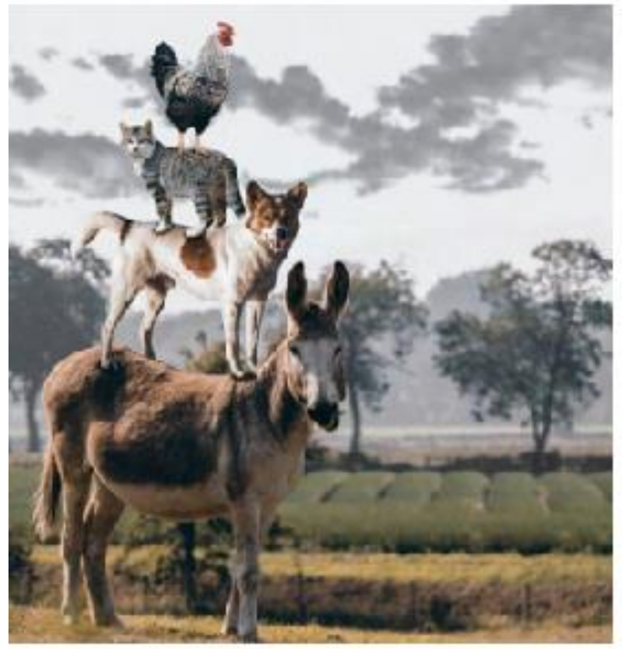
Highlight: **Zero-shot Multimodal Chain-of-Thought**



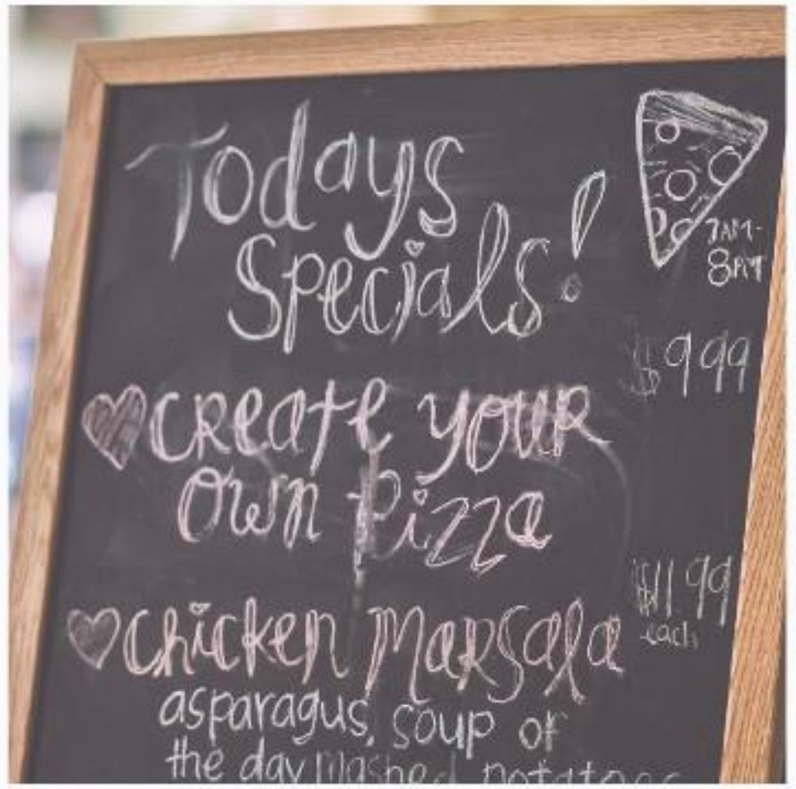
Given <img>.
Q: In this photo, are there more cats, or more dogs?
A: Let's think step by step.
There are 2 dogs and 2 cats.
2 and 2 are equal.
Therefore there are equal number of cats and dogs.



Given <img>. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step.
1. do not enter.
2. except bicycles.
3. do not entry except bicycles.
4. yes.



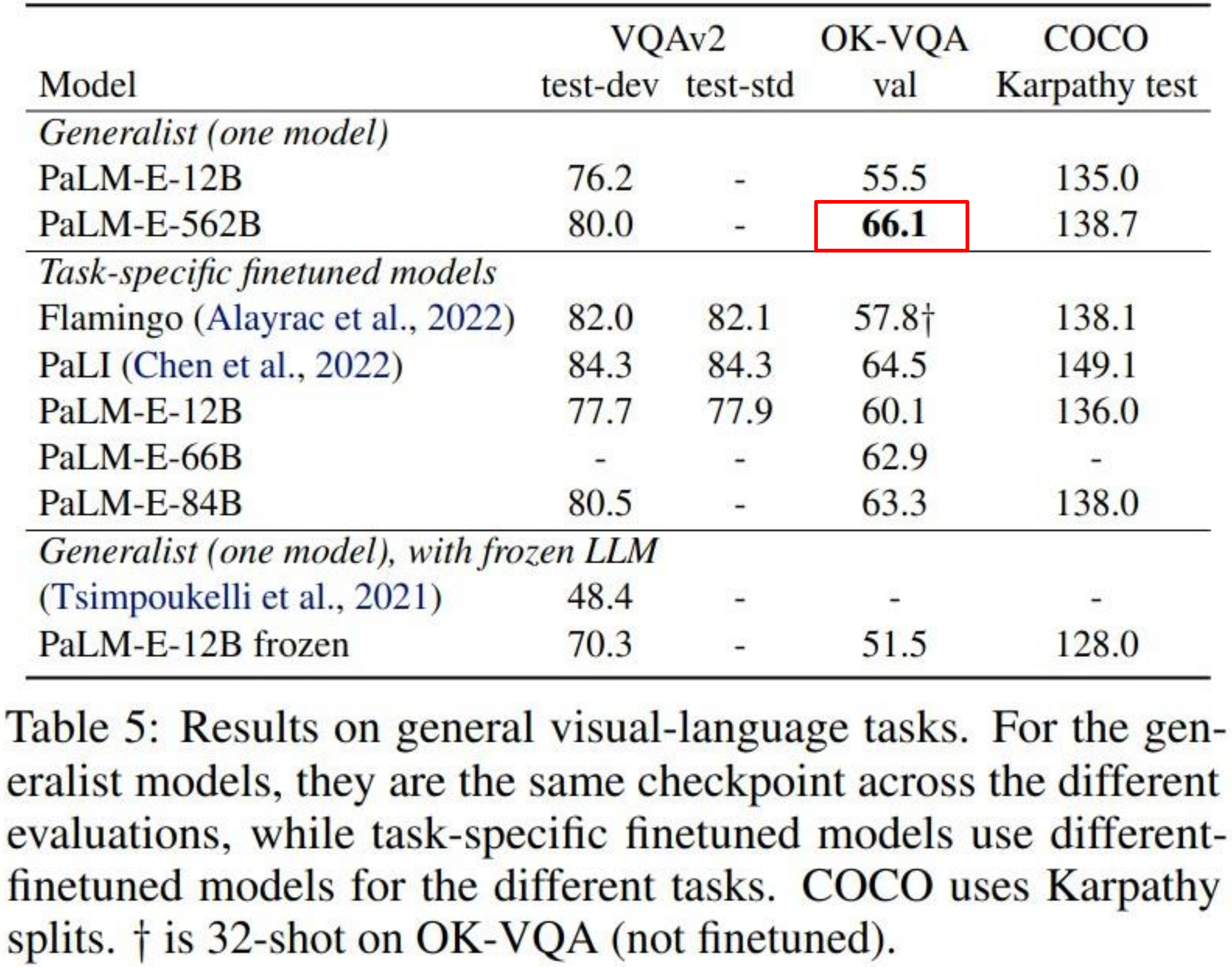
I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see <img>. Description: a donkey is carrying a dog, cat, and rooster. Joke: what do you call a donkey with a rooster on his back. A rooster booster.



Given <img> .
Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let's think step by step.
1st step: 2 custom pizzas.
2nd step: 2 x $9.99.
3rd step: $19.98.

Experiments and Results

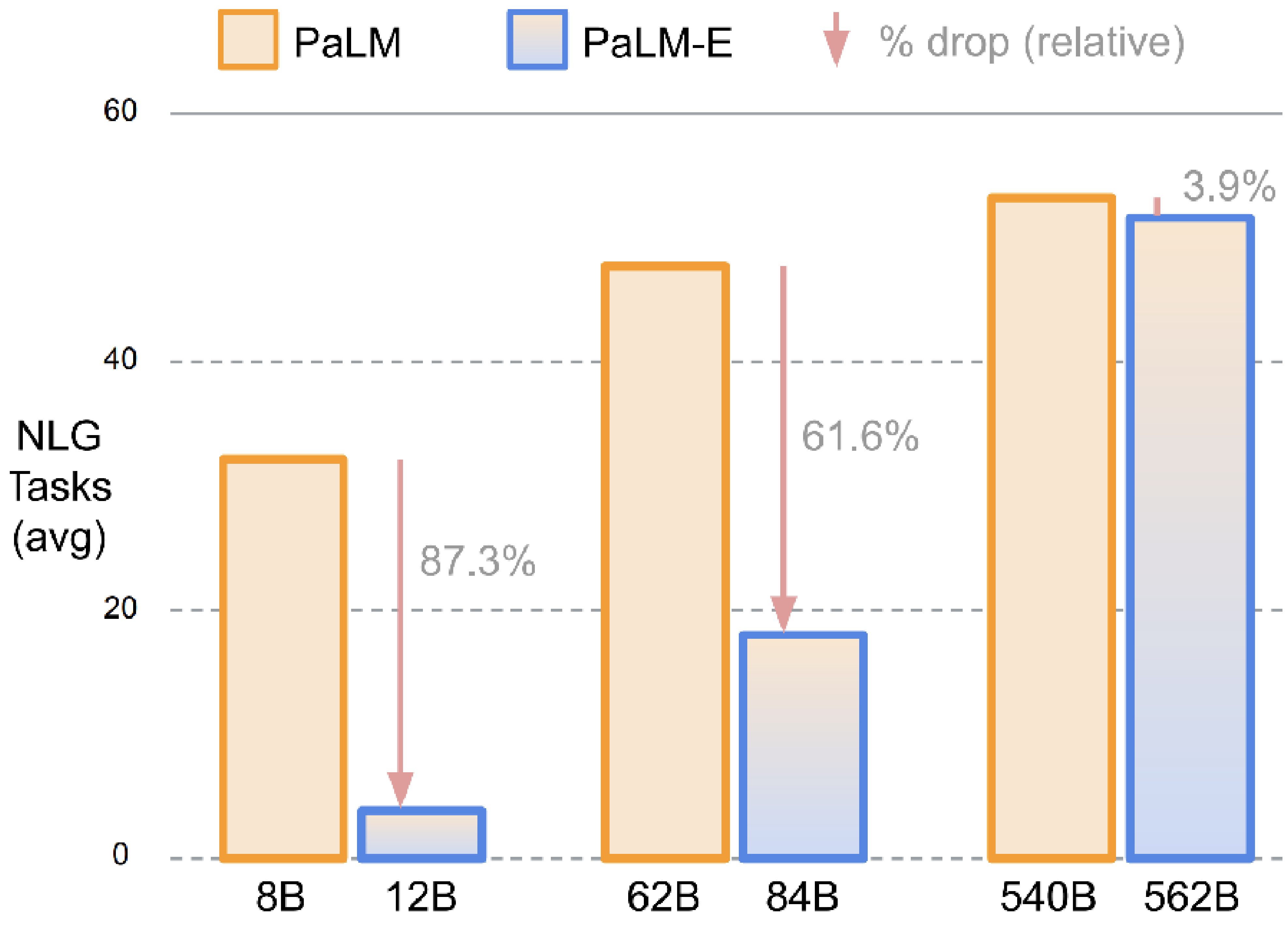
General Vision-Language Tasks



	VQAv2		OK-VQA	COCO
Model	test-dev	test-std	val	Karpathy test
Generalist (one model)				
PaLM-E-12B	76.2	-	55.5	135.0
PaLM-E-562B	80.0	-	**66.1**	138.7
Task-specific finetuned models				
Flamingo (Alayrac et al., 2022)	82.0	82.1	57.8†	138.1
PaLI (Chen et al., 2022)	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0
Generalist (one model), with frozen LLM				
(Tsimpoukelli et al., 2021)	48.4	-	-	-
PaLM-E-12B frozen	70.3	-	51.5	128.0

Table 5: Results on general visual-language tasks. For the generalist models, they are the same checkpoint across the different evaluations, while task-specific finetuned models use different-finetuned models for the different tasks. COCO uses Karpathy splits. † is 32-shot on OK-VQA (not finetuned).

Language catastrophic forgetting reduced with scale



Strengths and Weaknesses of PaLM-E

Strengths:

- Injects web-scale knowledge for diverse embodiments and tasks.
- Integrates real-world continuous sensor modalities into an LLM.
- Shows strong generalization and transfer capabilities.

Weaknesses:

- Performance is bottlenecked by low-level control (RT-1 model).
- Bias in LLMs used will propagate to the robotics tasks as well.
- Concatenating visual and linguistic features and applying cross-attention might fail to align the distinct statistical properties of pixel level and word-level representations.

PaLM-E: Pathways Language Model - Embodied

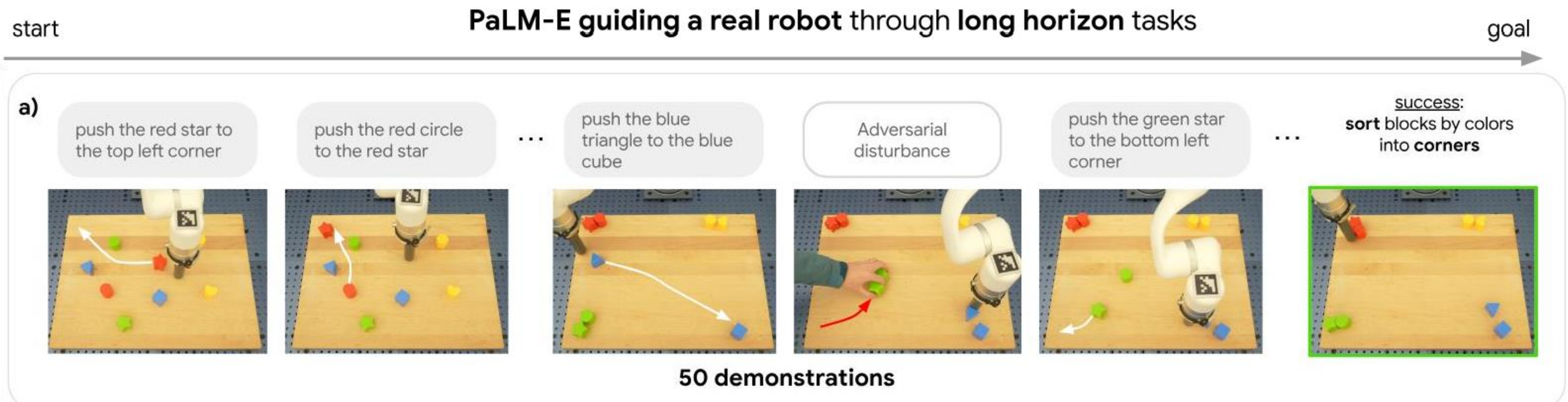
- The name of the model implies this is an embodied agent, however, is this actually the case or is it just good clickbaiting strategy?

- Agency: Learning from feedback and selecting outputs so as to pursue goals.

- Embodiment: Modelling input-output contingencies, including systematic effects, and using this model in perception or control.

Embodiment: Agent must model how its outputs affect the environment to distinguish itself from the environment in perception or to facilitate motor control.

- PaLM-E is not trained end-to-end, low-level actions are performed by RL-1.
- It just receives the goal + latest image and predicts the next set of actions, forgetting prior plans entirely, no memory of past actions.
- It's observing differences between images not understanding causation.


start
PaLM-E guiding a real robot through long horizon tasks
goal
a)
push the red star to the top left corner
push the red circle to the red star
...
push the blue triangle to the blue cube
Adversarial disturbance
push the green star to the bottom left corner
...
success: sort blocks by colors into corners
50 demonstrations

Agency: Learning from feedback and selecting outputs so as to pursue goals.

Both of the main components of the system are trained to imitate human behaviours:

- PaLM-E is trained to predict the next token in human-generated strings.
- The policy unit, RL-1, is trained to imitate human visuomotor control.

- PaLM-E is an autoregressive model, it has no feedback loop.

- It does not learn to pursue goals from feedback about success or failure.

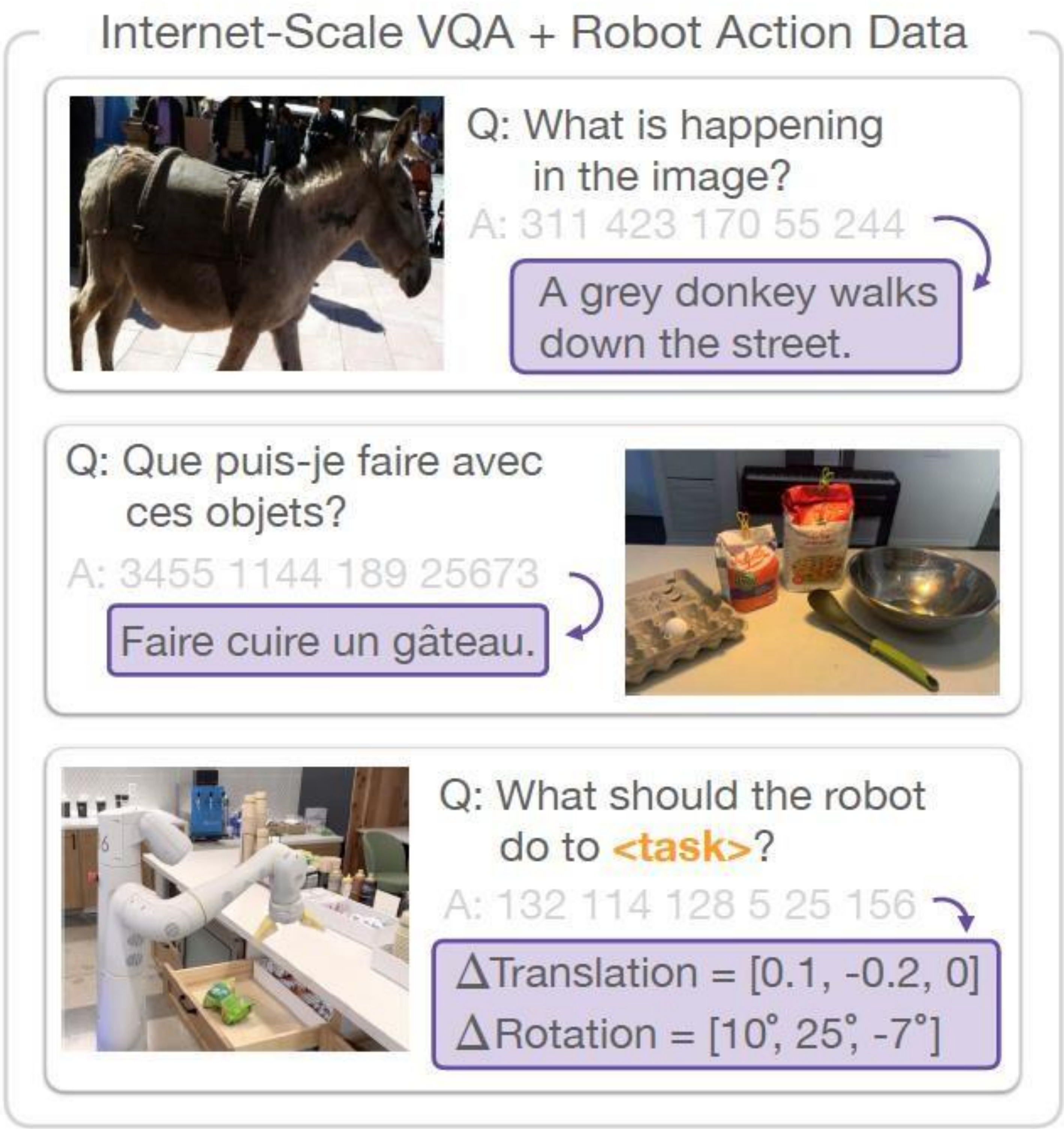
- It predicts a new action conditioned on the new image, NOT a revision of its earlier plan.

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

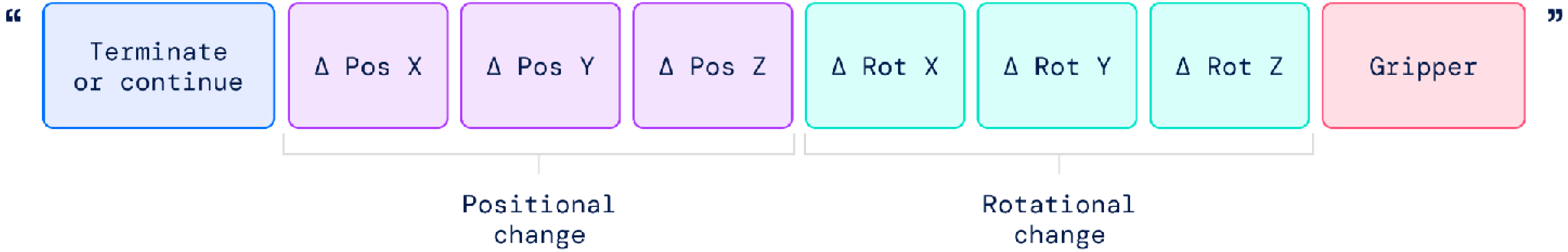
Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, Brianna Zitkovich

RT-2

- Details:
  - Use PaLM-E VLM backbone
  - Co-fine-tuning it on robotic data and VQA data
  - Input: (observation image, task description)
  - Output: action tokens (256)
- For PaLM-E, we need to overwrite 256 least used tokens in the vocabulary


Internet-Scale VQA + Robot Action Data
Q: What is happening in the image?
A: 311 423 170 55 244
A grey donkey walks down the street.
Q: Que puis-je faire avec ces objets?
A: 3455 1144 189 25673
Faire cuire un gâteau.
Q: What should the robot do to <task>?
A: 132 114 128 5 25 156
ΔTranslation = [0.1, -0.2, 0]
ΔRotation = [10°, 25°, -7°]

Representing Actions in VLMs



“ Terminate or continue | Δ Pos X | Δ Pos Y | Δ Pos Z | Δ Rot X | Δ Rot Y | Δ Rot Z | Gripper ”

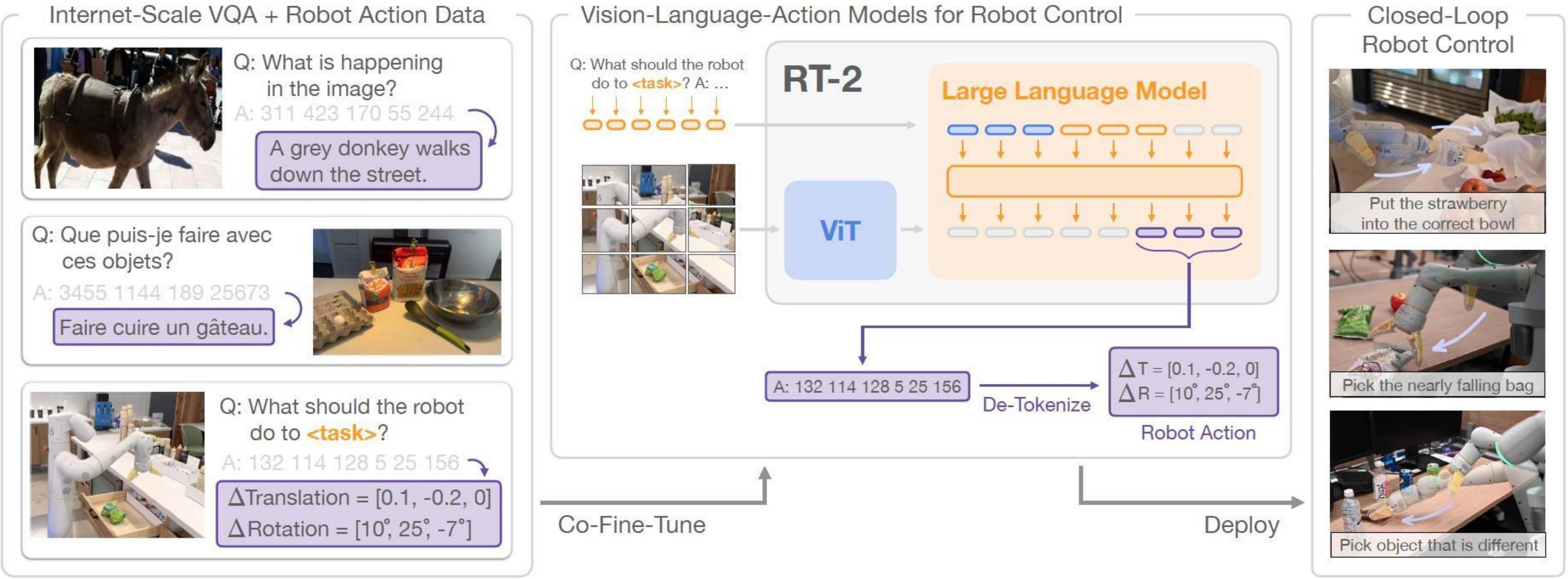
Positional change

Rotational change

- **Robot actions:**
  - Moving the robot arm and gripper
  - Discretized into 256 bins
- **Actions in VLMs**
  - Convert to a string of numbers
  - Example: “1 127 115 218 101 56 90 255”
  - Alternatives:
    - *Float numbers* - more tokens needed
    - *Human language (left, right etc.)* - can’t be directly executed on a robot

→ **Vision-Language-Action (VLA) model!**

RT-2



Internet-Scale VQA + Robot Action Data

Q: What is happening in the image?

A: 311 423 170 55 244

A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets?

A: 3455 1144 189 25673

Faire cuire un gâteau.

Q: What should the robot do to <task>?

A: 132 114 128 5 25 156

ΔTranslation = [0.1, -0.2, 0]

ΔRotation = [10°, 25°, -7°]

Vision-Language-Action Models for Robot Control

Q: What should the robot do to <task>? A: ...

RT-2

Large Language Model

ViT

A: 132 114 128 5 25 156

De-Tokenize

ΔT = [0.1, -0.2, 0]

ΔR = [10°, 25°, -7°]

Robot Action

Co-Fine-Tune

Deploy

Closed-Loop Robot Control

Put the strawberry into the correct bowl

Pick the nearly falling bag

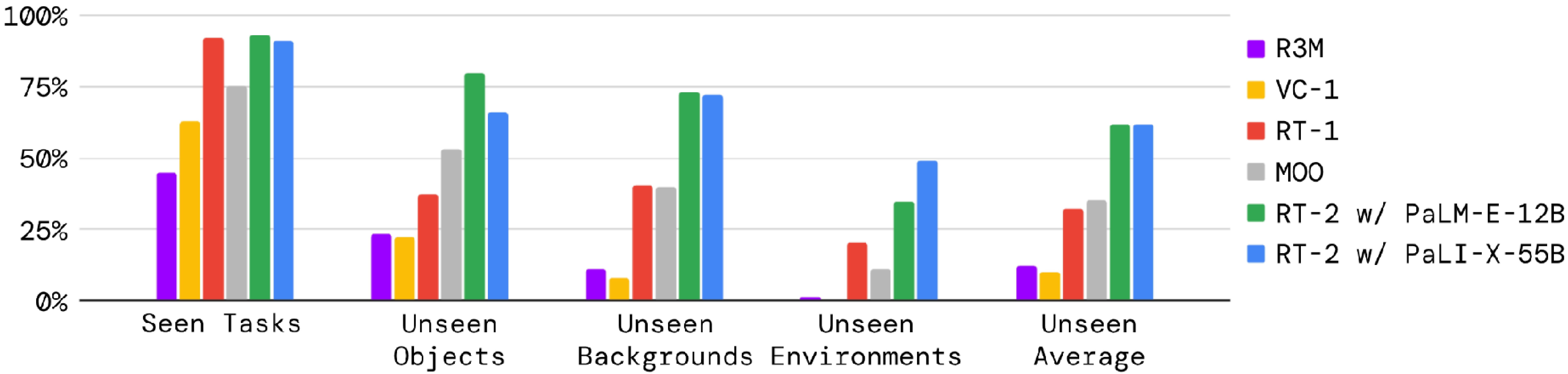
Pick object that is different

Results: Quantitative evals

(a) Unseen Objects

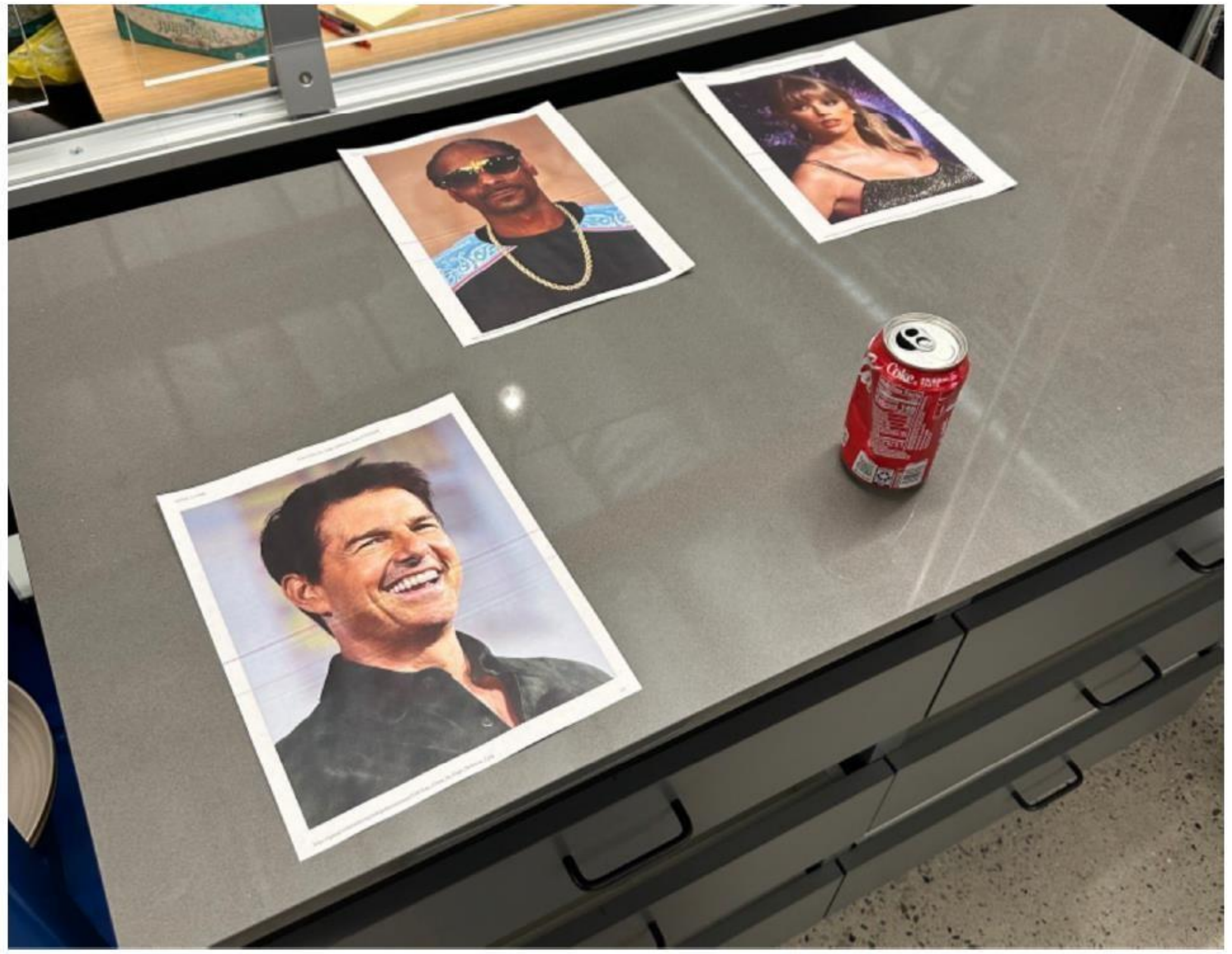
(b) Unseen Backgrounds

(c) Unseen Environments



RT-2

- It understand human instructions

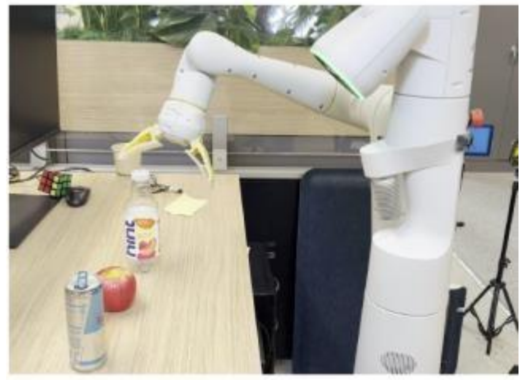


“move coke can to Taylor Swift”

User

I am sleepy, bring me a drink to help.

Plan: Pick redbull can

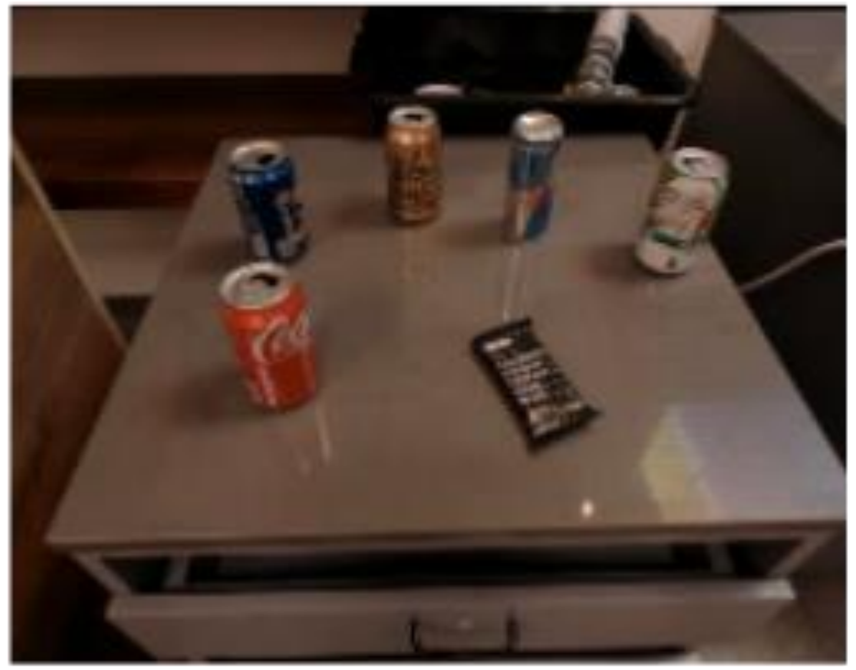


Results: Chain-of-Thought with RT-2-PaLM-E

Prompt:
Given <img> Instruction:
Bring me a drink.
Prediction:
Plan: pick 7up can.
Action: 1 143 129 123 145
114 115 127

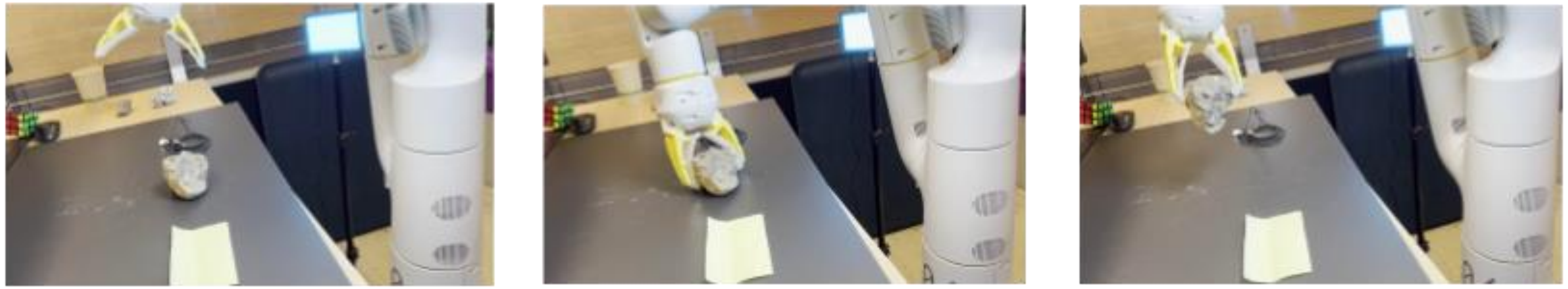
Prompt:
Given <img> Instruction:
Move all the objects
together.
Prediction:
Plan: move green can near
green rice chip bag.
Action: 1 128 126 127 135
123 119 127

Prompt:
Given <img> Instruction:
Pick the object that is
different from all other
objects
Prediction:
Plan: pick rxbar
chocolate. Action: 1 128
129 125 131 125 128 127



Prompt:
Given <img> Instruction:
Move the green objects
together.
Prediction:
Plan: move green can near
green rice chip bag.
Action: 1 130 129 121 131
127 128 127

Prompt:
Given <img> I need to
hammer a nail, what
object from the scene
might be useful?
Prediction:
Rocks. Action: 1 129 138
122 132 135 106 127



RT-2: Pros and Cons

- Pros:
  - No need for RT-1 intermediary.
  - Shows good scalability and transfer of knowledge across domains.
- Cons:
  - Slow to run, and difficult to deploy into new scenes and embodiments, needs fine-tuning.
  - It is still relying on a VLM as its backbone, which was not trained to output robotic actions and is therefore not truly grounded.
  - Concatenating visual and linguistic features and applying cross-attention might fail to align the distinct statistical properties of pixel level and word-level representations.

Bibliography

PaLM-E
Huang, W. *et al.* (2023). PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint* arXiv:2303.03378.

RT-1
Brohan, A. *et al.* (2022). RT-1: Robotics Transformer for Real-World Control at Scale. *arXiv preprint* arXiv:2212.06817.

RT-2
Brohan, A. *et al.* (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint* arXiv:2307.15818.

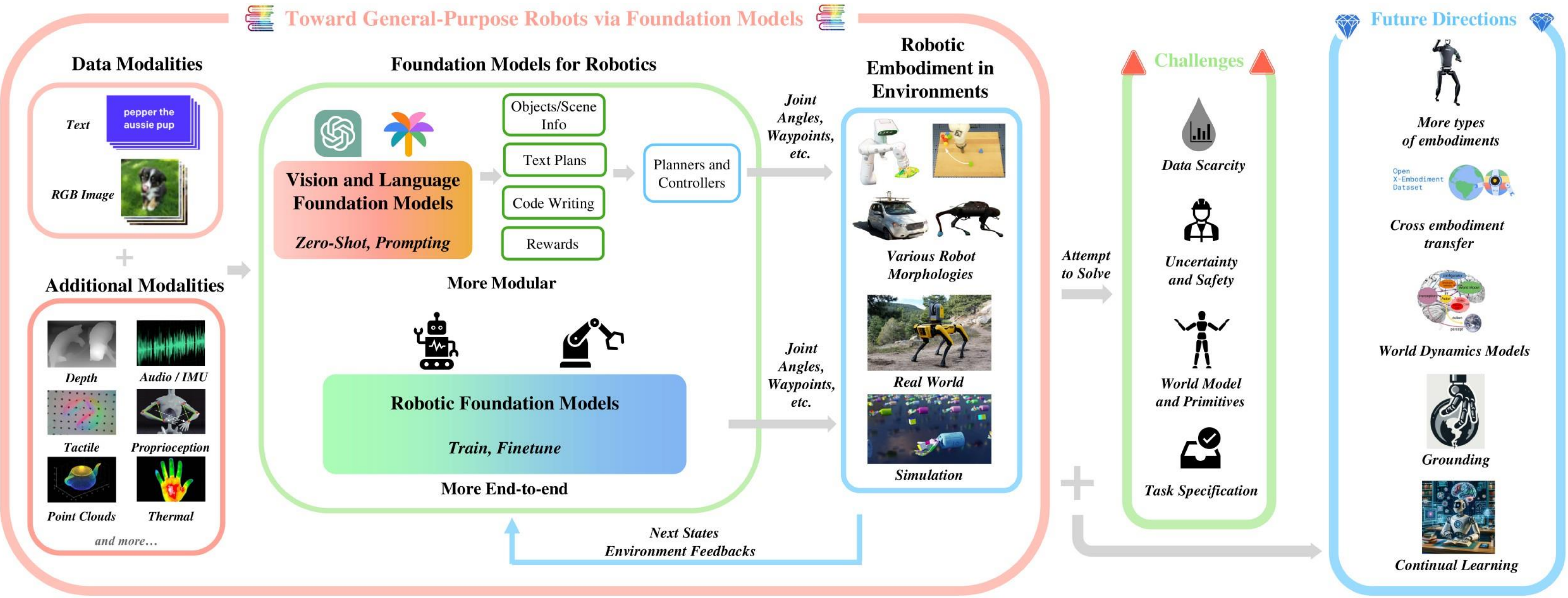
Vision-Language-Action review
Din, M. U., Akram, W., Saoud, L., Rosell, J., & Hussain, I. (2025). Vision Language Action Models in Robotic Manipulation: A Systematic Review. *arXiv preprint* arXiv:2507.10672.

Consciousness in AI
Butlin, P. *et al.* (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint* arXiv:2308.08708.

Georgia Tech CS7643 lecture
Xia, F. (2024). Lecture 21 slides for CS7643: Deep Learning. Georgia Institute of Technology. Available at: https://sites.cc.gatech.edu/classes/AY2024/cs7643_fall/assets/L21_Fei.pdf.

Thank you for listening!!!

Any questions?



Toward General-Purpose Robots via Foundation Models

Data Modalities

Text

pepper the aussie pup

RGB Image

+

Additional Modalities

Depth

Audio / IMU

Tactile

Proprioception

Point Clouds

Thermal

and more…

Foundation Models for Robotics

Vision and Language Foundation Models

Zero-Shot, Prompting

Objects/Scene Info

Text Plans

Code Writing

Rewards

Planners and Controllers

More Modular

Robotic Foundation Models

Train, Finetune

More End-to-end

Joint Angles, Waypoints, etc.

Joint Angles, Waypoints, etc.

Robotic Embodiment in Environments

Various Robot Morphologies

Real World

Simulation

Next States
Environment Feedbacks

Attempt to Solve

Challenges

Data Scarcity

Uncertainty and Safety

World Model and Primitives

Task Specification

+

Future Directions

More types of embodiments

Open X-Embodiment Dataset

Cross embodiment transfer

World Dynamics Models

Grounding

Continual Learning